RTO-MP-45
AC/323(SCI)TP/19

RTO-MP-45

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO MEETING PROCEEDINGS 45

Search and Target Acquisition

(Recherche et acquisition d'objectifs)

Papers presented at the RTO Workshop organised by the Systems Concepts and Integration Panel (SCI), held in Utrecht, The Netherlands, 21-23 June 1999.

Published March 2000

Distribution and Availability on Back Cover

RTO-MP-45
AC/323(SCI)TP/19

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO MEETING PROCEEDINGS 45

Search and Target Acquisition

(Recherche et acquisition d'objectifs)

Papers presented at the RTO Workshop organised by the Systems Concepts and Integration Panel (SCI), held in Utrecht, The Netherlands, 21-23 June 1999.

The Research and Technology Organization (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote cooperative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective coordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also coordinates RTO's cooperation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of initial cooperation.

The total spectrum of R&T activities is covered by 7 Panels, dealing with:

- SAS Studies, Analysis and Simulation
- SCI Systems Concepts and Integration
- SET Sensors and Electronics Technology
- IST Information Systems Technology
- AVT Applied Vehicle Technology
- HFM Human Factors and Medicine
- MSG Modelling and Simulation

These Panels are made up of national representatives as well as generally recognised 'world class' scientists. The Panels also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier cooperation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

Printed on recycled paper

Published March 2000

ISBN 92-837-1035-5

Printed by Canada Communication Group Inc.
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

Search and Target Acquisition

(RTO MP-45)

Executive Summary

Background. Standardized methods are needed to evaluate the effectiveness of camouflage, concealment and deception (CCD). CCD refers to all detection avoidance techniques, including netting, painted patterns, terrain cover, smoke and obscurants, shape-modifying applique, low-emissive paints, advanced contrast reduction materiel, and other technologies.

The SCI-12 Working Group was established to address this need. The scope was limited to defeating imaging, man-in-the-loop systems, specifically the unaided eye, direct view optics and electro-optical imagers. Non-imaging sensors and automatic target detection were not addressed. To facilitate the objective evaluation of alternative methodologies, researchers from the NATO countries were invited to apply their preferred methods on a standard set of 44 images of military vehicles in operational poses for which human observer search and target acquisition performance data were available. The results were reviewed at a June 1999 workshop in Utrecht, Netherlands.

The goal was to define measurement methods and signature metrics that are highly correlated with operational effectiveness, NOT to predict search time and probability of detection. Search time and probability of detection will be different in different military situations due to threat observer factors independent of the CCD signature (e.g., the relative penalty for missed detection versus false report, time pressure, fatigue, workload, familiarity with the terrain, prior expectations, etc.). Other criteria included low burden, repeatability, applicability during both design iteration and prototype evaluation, and robustness over CCD techniques, targets, and terrain.

Findings and Recommendations. At the present time, man-in-the-loop assessment is the only robust and effective method to evaluate CCD. Computational signature analysis methods are not sufficiently mature; they do not represent the range of significant visual and cognitive processes driving target acquisition performance.

The recommended approach is to evaluate target conspicuity using a standardized procedure to measure the ***visual lobe*** in off-axis detection, and ***response time*** in foveal examination. The visual lobe is the largest angle between the target and eye fixation at which the target can be discriminated. The visual lobe may be zero for hard-to-detect targets, and additional time is spent on foveal examination. Response time is measured for the entire test, then data analysis extracts the time spent on foveal examination.

Tests have shown that *(1)* target conspicuity is highly correlated with search time and probability of detection, and *(2)* target conspicuity measured in the laboratory is highly correlated with the corresponding measurements made in the field. Actual targets and sensors are used in the field. Photographic, synthetic or hybrid images with simulated sensor effects are used in the laboratory.

The procedure is simple and fast. It is stable with small number of observers. Minimal special equipment is needed. The procedure is well-suited to use during the design process to evaluate CCD alternatives, to support development and selection of robust and effective equipment, and to optimize CCD in the field.

Search and target acquisition field tests are orders of magnitude more costly and time consuming, yet provide assessment only for limited terrain and seasonal conditions. They require fielded equipment and can not support the early design process. Field test conditions are sufficiently unlike military operations that the results are not reliable predictors of effectiveness in real military operations.

Unresolved Issues. The committee did not produce standards or guidelines for *(1)* synthetic (simulated) images, *(2)* image capture in the field, *(3)* synthetic target insertion into captured images, or *(4)* image presentation in laboratory tests. These issues warrant further consideration.

Recherche et acquisition d'objectifs

(RTO MP-45)

Synthèse

Préambule. Des méthodes homologuées sont nécessaires pour évaluer l'efficacité des techniques de camouflage, dissimulation et déception (CCD). Le terme CCD englobe l'ensemble des techniques de prévention de la détection, comme les filets, les motifs peints, le masquage par le terrain, la fumée et les obscurcissants, les plaques de blindage dissimulant les formes réelles, les peintures à émissivité réduite, le matériel avancé de réduction de contraste et autres technologies.

Le groupe de travail SCI-12 a été créé pour répondre à ce besoin. Le mandat a été limité à la neutralisation des systèmes d'imagerie où l'homme est dans la boucle, et en particulier l'oeil nu, les optiques à vision directe et les systèmes à imagerie électro-optiques. Les capteurs sans imagerie et la détection automatique d'objectif n'ont pas été examinés. Afin de faciliter l'évaluation objective des méthodologies de remplacement, un certain nombre de chercheurs des pays membres de l'OTAN ont été invités à appliquer leurs méthodes préférées à un jeu standard de 44 images de véhicules militaires dans des configurations opérationnelles pour lesquels des données relatives aux performances d'observateurs humains en recherche et acquisition d'objectif étaient disponibles. Les résultats ont été étudiés lors d'un atelier organisé au mois de juin 1999 à Utrecht, aux Pays-Bas.

Cet atelier avait pour objectif de définir des méthodes de mesure et des paramètres de signature en corrélation étroite avec l'efficacité opérationnelle, et non de prévoir les temps de recherche et la probabilité de détection. La probabilité de détection et les temps de recherche sont différents selon la situation militaire, du fait de certains facteurs affectant les observateurs et indépendants de la signature CCD (par exemple la pénalisation relative liée à une détection manquée par rapport à une fausse indication, les contraintes temporelles, la charge de travail, la familiarité avec le terrain, les attentes etc.) Parmi les autres critères pris en compte il faut citer la facilité d'emploi, la capacité de répéter les opérations très rapidement, la capacité de les appliquer durant les itérations des étapes de conception et durant l'évaluation des prototypes, ainsi que la robustesse face au CCD ennemi, aux types d'objectif et aux types de terrain.

Conclusions et recommandations. Pour l'instant l'évaluation de type « homme dans la boucle » est la seule méthode sûre et efficace pour l'évaluation du CCD. Les méthodes informatisées d'analyse de la signature ne sont pas encore suffisamment au point; elles ne couvrent pas encore suffisamment l'éventail des processus visuels et cognitifs qui déterminent les performances en acquisition d'objectif.

L'approche préconisée est donc d'évaluer la perceptibilité de l'objectif suivant une procédure homologuée de mesure du ***lobe visuel*** lors d'une détection excentrée, et du ***temps de réponse*** pour l'examen fovéal. Le lobe visuel est le plus grand angle entre l'axe oeil-objectif et la direction du regard à laquelle il est possible de faire la discrimination de cet objectif. Le lobe visuel peut être de zéro pour les objectifs difficiles à détecter, et une période supplémentaire est alors nécessaire pour l'examen fovéal. Le temps de réponse est mesuré pour toute la durée de l'essai, puis le temps passé à l'examen fovéal est extrait par analyse des données.

Les essais ont permis de démontrer que *(1)* il existe une forte corrélation entre la perceptibilité de l'objectif et le temps de recherche, ainsi qu'avec la probabilité de détection, et *(2)* qu'il existe également une forte corrélation entre la perceptibilité de l'objectif mesurée en laboratoire et les mesures correspondantes effectuées sur le terrain. Des objectifs et des capteurs réels ont été utilisés sur le terrain. Des images photographiques, synthétiques ou hybrides, avec des effets simulés de capteurs ont été utilisés en laboratoire.

La procédure est simple et rapide. Elle est stable avec un nombre réduit d'observateurs. Elle nécessite un minimum de matériel spécialisé. La procédure est bien adaptée aussi bien pour l'évaluation de variantes de CCD lors du processus de conception, que pour le développement et le choix d'équipements robustes et efficaces et pour l'optimisation du CCD sur le terrain.

Les essais sur le terrain de la recherche et l'acquisition d'objectif sont beaucoup plus longs et coûteux, et ne donnent des évaluations que pour un nombre limité de terrains et de conditions saisonnières. Ils nécessitent un matériel qui doit être déployé sur le terrain et ne sont pas applicables lors des phases initiales de la conception. Enfin, les conditions dans lesquelles les essais sur le terrain sont effectués sont suffisamment différentes de celles des opérations militaires pour disqualifier les résultats en tant qu'indicateurs fiables d'efficacité en situation réelle.

Questions en suspend. Le comité n'a pas fourni de normes ni de directives en matière *(1)* d'images synthétiques (simulées), *(2)* d'images captées sur le terrain, *(3)* d'incrustation d'objectifs synthétiques dans des images saisies, ou *(4)* de présentation d'images pour essais en laboratoire. Ces questions mériteraient donc un examen ultérieur.

Contents

SESSION II: TARGET ACQUISITION MECHANISMS
Chairman: Ian Moorhead (UK)

SESSION III: SIMULATION ISSUES
Chairman: Mark Rodgers (UK)

† Paper was not presented at the Workshop.

Theme

The focus of the Workshop is on methods to evaluate the effectiveness of signature reduction countermeasures such as Camouflage, Concealment, and Deception (CCD). There are two different main approaches to evaluating CCD: (1) *observer experiments* and (2) *computational methods* using image analysis. While such analyses may extend to both imaging and non-imaging systems with and without a man in the loop, this workshop focusses on image-forming systems in which a *human* provides the primary *information processing*. The efforts concentrate on the visual and electro-optical signatures of the targets and their associated CCD treatments.

TOPICS

The following topics are addressed:

- methods to evaluate the effectiveness of CCD to increase survivability;
- search and target acquisition models incorporating CCD;
- computational target signature metrics;
- the design and evaluation of CCD equipment and techniques.

Acknowledgements

The Systems Concepts and Integration Panel wishes to express its thanks to the National Authorities of The Netherlands for the invitation to hold this workshop in their country.

We would like to thank the TNO Human Factors Research Institute for hosting the workshop and for providing the necessary logistic support.

And finally, we wish to thank the following for their contribution to the success of this workshop:

Turing Associates, Inc. (USA)
United States Air Force European Office of Aerospace Research and Development
United States Army Research Development and Standardization Group (UK)
United States Office of Naval Research, Europe
Signaal (NL)
TNO-HFRI (NL)

Systems Concepts and Integration Panel

CHAIRMAN:

Dr. Edwin B. STEAR
Eaton Hill Systems and Technologies
2103 Hunters Crest Way
Vienna, VA 22181
United States

VICE-CHAIRMAN:

Prof. Luis M.B. da Costa CAMPOS
Instituto Superior Tecnico
Torre-Norte 6.24
Avenida Rovisco Pais
1049-001 Lisboa Codex
Portugal

SCI-012 CCD Evaluation Techniques Chairman

Mr Randall R. WILLIAMS
USAE-WES, 3909 Halls Ferry Road
ATTN: CEWES-SS-C
Vicksburg, MS 39180-6199, USA

SCI-045 Programme Committee Chairman:

Dr Alexander TOET
TNO Human Factors Research Institute
Kampweg, 5, 3769 DE Soesterberg
The Netherlands

Members

Prof. Marc ACHEROY
Royal Military Academy
Avenue de la Renaissance
1040 Brussels
Belgium

Ms Ann BATCHELOR
TRACOR Aerospace
6500 Tracor Lane, MS 24-1
Austin, TX 78666, USA

Mr Christian M. BIRKEMARK
DDRE
P.O. Box 2715
Ryvangs Alle, 1
DK-2100 Copenhagen Ø
Denmark

Mr William CAPLAN
NATO C3 Agency
P.O.Box 174
2501 CD The Hague
The Netherlands

Dr Jean DUMAS
DREV
2459 Boul. Pie XI North
Val Belair, Quebec G3J 1X5
Canada

Mr Grant GERHART
US Army Tank-Automotive and
Armaments Command
Res., Development and Eng. Center
Warren, MI 48397-5000, USA

Mr Eddie JACOBS
NVESD, AMSEL-RD-NV-ST-VMS
10221 Burbeck Road, Suite 430
Fort Belvoir, VA 22060-5806, USA

Mr Steve LUKER
AFRL/VSBE
29, Randolph Road
Hanscom AFB, MA 01731, USA

Dr Thomas J. MEITZLER, Ph.D
TACOM, AMSTA-TR-R MS263
Vehicle Detection Team
Warren, MI 48397-5000, USA

Dr Kevin MITCHELL
Defence Clothing & Textiles Agency
S&T Division
Flagstaff Road, Colchester
Essex CO2 7SS
United Kingdom

Dr Ian R. MOORHEAD
Protection and Performance Dept.
Centre for Human Sciences
DERA
Fort Halstead, Sevenoaks
Kent, TN14 7BP
United Kingdom

Mr Mark L.B. RODGERS
Defence Clothing & Textiles Agency
S&T Division
Flagstaff Road, Colchester
Essex CO2 7SS
United Kingdom

Mrs Lucille SCHRADER
DCE/ETAS, B.P. 36
49460 Montreuil-Juigne
France

Mr Randy K. SCOGGINS
Waterways Experiment Station
ATTN: CEWES-SS-C
3909 Halls Ferry Road
Vicksburg, MS 39180, USA

Dr Mathee VALETON
TNO Human Factors Research
Institute
Kampweg, 5
3769 DE Soesterberg
The Netherlands

Mr Gary WITUS
Turing Associates, Inc.
1392 Honey Run Drive
Ann Arbor, MI 48103, USA

PANEL EXECUTIVE

LTC Scott CAMPBELL, USA

From USA:
RTA/SCI
PSC 116
APO, AE 09777

From other countries:
RTA/SCI
BP 25
7, rue Ancelle
92201 Neuilly sur Seine Cedex
France

TECHNICAL EVALUATION REPORT

Alexander Toet
TNO Human Factors Research Institute
Kampweg 5, 3769 DE Soesterberg, The Netherlands
Email: toet@tm.tno.nl

INTRODUCTION

The Workshop on Search & Target Acquisition was initiated by the Systems Concepts and Integration Panel SCI-12 (the former RSG-2), on "Camouflage, Concealment and Deception Evaluation Techniques". The goal of this workshop was to provide a state of the art review of CCD evaluation methodologies. In particular:

- to provide a forum for exchange of ideas,
- to compare current methodologies on standard data sets,
- to establish metrics for comparison of methods,
- to allow interaction between users and developers,
- to identify new directions for future assessment,
- methods and research programmes.

The main topics identified for the workshop can be summarised as:

- methods to evaluate the effectiveness of CCD;
- search and target acquisition models incorporating CCD;
- computational target signature metrics;
- the design and evaluation of CCD equipment and techniques.

To facilitate the objective evaluation of alternative CCD evaluation methodologies, researchers from the NATO countries were invited to apply their preferred methods on a standard set of 44 high resolution digitised images[1] of military vehicles in operational poses for which human observer search and target acquisition performance data were available.

OPENING ADDRESS

Cdre Ir. D. van Dord of the Dutch Ministry of Defense gave an Opening Address (not reprinted here), in which he outlined the importance of survivability enhancement through acquisition avoidance, including methods to evaluate the effectiveness of signature countermeasures such as CCD. He emphasized the need for standardised CCD evaluation techniques.

KEYNOTE ADDRESS

In her Keynote Address, entitled "Perception Studies", Barbara L. O'Kane of the Night Vision & Electronic Sensors Directorate, USA, discussed some challenges involved in perception studies that are conducted to gain insight into surveillance and target acquisition by military users of thermal imagery. On today's battlefield ground-to-ground and air-to-ground military target acquisition is performed with thermal sights. It is extremely important, therefore, to correctly train, model, and understand the use of these systems to prevent fratricide and increase survivability. It is not possible to perform all the needed research by means of expensive field tests. Therefore, the use of perception studies has become popular for developing training, testing system designs, and assessing effectiveness of camouflage techniques. For the outcome to be valid, a perception study should emulate as accurately as possible what a military observer will actually see and do when using a thermal sight. Dr. O'Kane identified and discussed five general issues: training, field-of-regard versus field-of-view search, gain and level controls, time limitations, and subject motivation. She argued that, in order to make the link between the laboratory perception study and the battlefield experience most robust, the methodology that is chosen should be directly related to a military operational scenario. She concluded that as technology allows greater capability to provide optimal emulation of military operational procedures, perception studies can become more and more realistic, which is important for developers and users of camouflage in evaluation of systems.

SESSION I: SEARCH PERFORMANCE PREDICTIONS

The thirteen papers in this session, that was chaired by Gary Witus of Turing Associates, Inc., USA, are concerned with issues in computational techniques to predict the human visual search and detection performance.

The first paper of this session, "Lessons Learned in Developing and Validating Models of Visual Search and Target Acquisition", by T.J. Doll of the Georgia Tech Research Institute, USA, and R. Home of the Defence and Evaluation Research Agency, UK, was presented by Ted Doll. He argued that complex pattern perception, visual attention, learning, and cognition are important factors in human visual search and target acquisition performance. He explained the contribution of these processes and suggested approaches for modeling them. He distinguished three different approaches for testing and validating models of visual search and target acquisition, that take into account the abovementioned factors. The first and most obvious approach is to compare model predictions to observer performance in the field. However, field experiments are expensive, time-consuming, and difficult to perform (environmental conditions are not under control, and unintended detection cues may be abundant). The second approach is to compare model predictions to observer data that is collected in the laboratory using imagery registered in the field (this approach was also addressed in the previous Keynote paper). The advantage of this method is that the observers and models are both subjected to the same image degradation effects. However, the deployment of real targets in the field is still expensive and time-consuming and may produce extraneous cues such as vehicle tracks. The third approach is to gather high resolution imagery of natural backgrounds, together with ground truth and data on meteorological and illumination conditions. Synthetic targets can then be generated and inserted in the calibrated background imagery, and the result validated by comparing it to field imagery (see Session III for papers on this topic). The synthetic images can then be used in laboratory observer experiments and the results can be compared to model

predictions. This method eliminates the disadvantages of the first two methods.

The second paper, "Visual Distinctness Determined By Partially Invariant Feautures", by J.A. Garcia, J. Fdez-Valdivia, X.R. Fdez-Vidal and R. Rodriguez-Sanchez of the University of Granada, Spain, was presented by Jose Garcia. He presented an algorithm for the automatically learned partitioning of "visual patterns" in digital images. The method is based on band-pass filtering operation, with fixed scale and orientation sensitivity. The "visual patterns" are defined as the features which have the highest degree of alignment in the statistical structure across different frequency bands. From this image representation he derived a computational visual distinctness measure. The measure was applied to quantify the visual distinctness of targets in the SEARCH_2 image database. The computed visual target distinctness measure correlates with the visual search and detection performance of human observers.

The third paper, "Target Detection Using Saliency-Based Attention", by L. Itti and C. Koch of the California Institute of Technology, USA, was presented by Laurent Itti. He presented a computer model of human visual search, based on the concept of a "saliency map", that is, an explicit two-dimensional map that encodes the saliency or conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location. He gave a detailed description of a computer implementation of this scheme, focusing on the problem of combining information across modalities, here orientation, intensity and color information, in a purely stimulus-driven manner. He showed examples of the successful application of this model to a wide range of target detection tasks, using synthetic and natural stimuli. He also presented predicted search times of his model on the SEARCH_2 image database of rural scenes containing a military vehicle. Overall, he found a poor correlation between human and model search times. Further analysis however revealed that in 3/4 of the images, the model appeared to detect the target faster than humans. It hence seems that this model, which was originally designed not to find small, hidden military vehicles, but rather to find the few most obviously conspicuous objects in an image, performed as an efficient target detector on the SEARCH_2 image dataset.

The fourth paper, "Applying the Law of Comparative Judgement to Target Signature Evaluation", by J.R. McManamey of the Night Vision and Electronic Sensors Directorate, USA, was presented by Eddie Jacobs, also of NVESD, USA. The Law of Comparative Judgement (LCJ) is a psychophysical tool that can be used to scale complex phenomena that lack easily identified physical parameters, such as target signatures. In a demonstration exercise, the author applied the LCJ to obtain a "search difficulty" value for a subset of the SEARCH_2 images. These LCJ scale values were compared to search times and probabilities of detection from a laboratory search experiment with human observers, performed by TNO-HFRI in the Netherlands. The scale values were not linearly related to search time and probability of detection, but correlated very well with the logarithm of mean search time ($r = 0.936$) and the cube of the number of correct responses ($r = 0.954$). A chi-squared goodness-of-fit test gave 94.6% confidence in the fit of the LCJ scale to the experimental data. While the LCJ results in a scale with no natural zero point and arbitrary units, this tool can be used to construct a standard scale. The author illustrated how a standard clutter scale might be constructed using the LCJ. He argued that the LCJ may be a useful tool in target signature evaluation, either when used in conjunction with scaling equations that permit conversion to familiar quantities such as mean search time and probability of detection, by providing relative "search difficulty" values, or by making possible a psychophysically meaningful clutter scale.

The fifth paper, "CAMEVA, A Methodology for Estimation of Target Detectability", was presented by the author, Christian Birkemark of the Danish Defence Research Establishment, Denmark. CAMEVA is a methodology for computerised CAMouflage EVAluation and for estimation of target detectability. Input is a single digitised image comprising a highly resolved target as well as a proper amount of background. Separate target and background images can also be handled. Target and background regions are manually selected. From the input data, CAMEVA predicts the target detectability as a function of the target distance. The detectability estimate is based on the dissimilarity of the statistical distributions of the target and background features. The extracted features should resemble those applied during the human perception process. Typically, contrast and various measures of edge strength are applied. The Bhattacharyaa distance establishes a relative separability, while the absolute detection range is obtained by deriving a relation between the Bhattacharyaa distance and the estimated target resolution, at range. By introducing parameters of the sensor, typically the human unaided eye, detectability as a function of the range is obtained. The methodology does not reflect individual observer performance, but provides an estimate of the optimal detection performance, given the selected set of features. CAMEVA depends strongly on the skills of the operator during the selection of target and background regions. The author considers to produce a kind of catalogue that will set up typical scenarios together with proposed operator methodologies to cope with these. He also plans to implement a proper procedure for modeling of atmospheric transmission loss and of light conditions. Fundamentals for these sub-models have been investigated, but still need validation. He also argued that the current choice of features is not necessarily optimal. Certain aspects of detection are currently not modeled. A typical example is the cueing provided by long straight lines. Further features need to be investigated and in some cases algorithms for their implementation must be developed. For the SEARCH_1 and 2 image datasets, the detection probablility estimated by CAMEVA correlates only weakly with observer performance.

The sixth paper, "Evaluation of Target Acquisition Difficulty Using Recognition Distance to Measure Required Retinal Area", was presented by the author, Thomy Nilsson, of the University of Prince Edward Island, Canada. He applied the method of limits to measure recognition distance thresholds for the vehicles in the SEARCH_2 images, both for calibrated slides produced from the digital imagery, and for the images presesented four times enlarged on a CRT. His rationale is that less visible targets should require more visual pathways for recognition, and that difficulty of acquisition can therefore be defined in terms of the relative retinal area required for recognition. He derived the relative retinal area from the inverse square of the recognition distance of a particular vehicle relative to the distance of the vehicle that could be seen furthest away. He compared the results with the mean search times for the vehicles in these pictures, determined by TNO-HFRI in The Netherlands. Both recognition distance thresholds and retinal difficulty correlated only weakly with mean search times. Analysis of the results showed that there is a significant effect of target size on retinal difficulty and recognition distance. He found no significant effect of target

contrast and shape on mean search time. He concluded that mean search time and recognition distance may be complementary measures of visual target distinctness.

The seventh paper, "Evaluating TNO Human Target Detection Experimental Results Agreement with Various Image Metrics", by G. Aviram and S. R. Rotman of the Ben-Gurion University of the Negev , Israel, was presented by Stan Rotman. The authors tested the agreement between the TNO-HFRI laboratory observer results on the SEARCH_2 imagery and four different target distinctness metrics, originally designed to evaluate detection performance of infrared imagery. The metrics they tested include two local target from background distinctness metrics (*DOYLE* and *TARGET*), a global image complexity metric (*POE*) and a textural global / local co-occurrence matrix metric (*ICOM*). Applying these metrics to the image database they obtained the highest correlation values between the experimental results and the two local metrics (*DOYLE* and *TARGET*) values, and somewhat lower correlation levels between the *ICOM* global / local texture metric values and the experimental results, and a very low correlation level between the *POE* global clutter metric values and the experimental results. However, none of the correlation levels exceeded 0.6. The authors conclude that

- *local target to background distinctness determines detection performance,*
- for targets with low visual distinctness, the global clutter level determines the detection performance.

They suggest to use these findings to define fuzzy-type classification rules. In their analysis they excluded images containing very large size targets (for evaluation of all the metrics), or very narrow extent targets (for evaluation of the *ICOM* metric).

The eighth paper, "Image Based Contrast-to-Clutter Modeling of Detection", was presented by the author, David L. Wilson of the Night Vision & Electronic Sensors Directorate, USA. He applied a range of image-based contrast metrics to calculate the visual distinctness of the targets in the SEARCH_2 images. The metrics included different combinations of the variance of the pixel values over the target and its local or estimated background, the difference between the mean pixel values of the target and its local background, the target size, defined as the number of pixels on target, and a modified version of the Schmieder-Weathersby clutter metric, computed either over the entire image or over a user defined region. He then correlated the predicted target distinctness values with the mean search time provided with the SEARCH_2 image dataset. His results show that a simple root mean square difference of the pixel values over the target and its local background area correlates most strongly with observer performance. The use of a local clutter metric does not improve this result, whereas the use of a global clutter metric lowers the correlation values. This finding clearly agrees with the results of the previous presentation, i.e. that local target distinctness appears to determine detectability. He concludes that the inclusion of a clutter metric in a target distinctness measure should in theory have advantages when there is a large variation in clutter, but that it was not possible to demonstrate this advantage with the SEARCH_2 dataset.

The nineth paper in this session, "Efficient Methods for Validating Target Acquisition Models", was presented by the author, Richard Hecker of IABG, Germany. He addressed the validation of the CAMAELEON computer model. This model is developed for the *assessment of camouflage* using digital image processing techniques based in part on the human visual system. It estimates the physiological *detectability* of an object by calculating the similarity between the object and its background relating to first order statistic features like *contrast* and textural features like *local contrast (energy), local spatial frequency and local orientation*. These local textural features are calculated from the output of several bandpass-filters, that are similar to the filters constituted by the receptive fields of the neurons in the early stages of the human visual system. The histograms of these local features are then calculated both for the object and its background. The overlap of the histograms of the target and its background is taken as a measure of their similarity. These similarity measures are combined in a heuristical detection model to calculate *(a)* the detectability probability as a function of range and *(b)* the detectability range. The model was validated with direct measurements of target detectability ranges in the field, both for infrared and visual imagery. For the SEARCH_2 dataset the detectability range predicted by CAMAELEON correlates with both the detection and identification lobes. The author argues that this correlation is a direct result of the variation in target sizes (viewing range) in the SEARCH_2 dataset, and that the interfering effects with different cues (size, atmosphere, resolution, contrast, texture) that arise from this variation in target size cannot be resolved. CAMAELEON is designed to analyse high resolution images taken from nearby, that are not degraded by atmospheric effects.

The tenth paper, "Assessing Camouflage Methods Using Textural Features", by S. Nyberg and K. Schutte, of respectively the Defence Research Establishment, Sweden and TNO-FEL, The Netherlands, was presented by Sten Nyberg. The authors applied a large range of local digital visual target distinctness measures to the SEARCH_2 imagery, and calculated the correlation of the computed values with the mean search time provided with the dataset. They found that local mean and variance based measures, together with directional autocorrelation and isotropy (both features that are based on the local power spectrum), yield correlations up to 0.85. They argue that isotropy works well because it reacts to small straight edge segments that are typical for targets but not characteristic for the background.

The eleventh paper, "Image Discrimination Models for Object Detection in Natural Backgrounds", was presented by the author, Al J. Ahumada Jr. of NASA Ames Research Center, USA. He applied a simple linear Difference-of-Gaussians filter model to small target sections of the SEARCH_2 image pairs, representing the same scene both with and without the target present. He calculated the visual target distinctness as the Euclidian distance between the filtered image pairs, normalised by the root-mean-square contrast of the filtered background-only image, including a global contrast threshold to emulate masking in the human visual system. This measure effectively combines *(1)* target size, *(2)* target contrast, and *(3)* local background contrast variability. The author obtained a correlation of 0.81 between the distinctness values thus calculated and mean search time provided for the SEARCH_2 imagery. This relatively high correlation is found despite the fact that the model does not account for *(1)* color differences, *(2)* target position, *(3)* object contours, or *(4)* texture differences.

The twelfth paper, "A Contrast Metric for 3-D Vehicles in Natural Lighting ", by G. Witus of Turing Associates, Inc, USA, and G. Gerhart of U.S. Army Tank-automotive and Armaments Command, USA, was presented by the author and chairman of this session, Gary Witus. He argued that basic vision research suggests that shape from shading and 3-D appearance are pop-out cues, focus visual attention, and facilitate figure-ground segregation. Although it works for stylized 2-dimensional targets, the standard area-weighted average contrast ratio has proven *not* to be a good predictor of

search and target acquisition performance for complex targets in complex scenes. The authors introduced a simple 3-D target contrast metric, that effectively combines the area weighted contrast over *(1)* the front/rear, *(2)* the top, and *(3)* the side regions of the projected view of a target vehicle. They applied this metric to the targets in the SEARCH_2 images. Their results show that this 3-D structure contrast metric performs better than RSS contrast (the Root-Sum Square of the target-background luminance difference and the target luminance standard deviation, which has been found to be an effective metric in previous studies), and both perform dramatically better than the area-weighted average contrast. Target height performs better than either target area or square root of area. The 3-D signature metric accounts for over 80% of the variance in probability of detection, and 75% of the variance in search time as measured in the TNO perception tests. When false alarm effects are discounted, the metric accounts for 89% of the variance in probability of detection and 95% of the variance in search time. The predictive power of the signature metric, when it is calibrated to half the data and evaluated against the other half, is 90% of the explanatory power. False alarms are a significant factor contributing to variance in search performance. The authors conclude that further research should address effective models to predict the rate of false alarm from image properties and top-down knowledge.

The thirteenth paper in this session, "Computing Search Time in Visual Images Using the Fuzzy Logic Approach by T.J. Meitzler, E. Sohn, H. Singh, and A. Elgarhi, of the US Army Tank-automotive and Armaments Command Research, and of the Wayne State University, USA, was not presented. This paper describes a fuzzy logic model that predicts mean search time from local luminance, range, aspect, width, and wavelet edge points. The authors claim to obtain a correlation of 0.97 between the model predictions and the mean search times provided for the SEARCH_2 image data set.

SESSION II: TARGET ACQUISITION MECHANISMS

The seven papers in this session, that was chaired by Ian Moorhead of the Defence Evaluation and Research Agency, UK, address psychophysical studies and computer models of different aspects of the human visual target acquisition capability.

The first paper in this session, "The Sources of Variability in the Search Process", was presented by the author, Kevin Cooke of the British Aerospace Sowerby Research Centre, UK. He discussed the considerations and parameter sensitivity analysis involved in the design of the statistical ORACLE model. ORACLE models the probability of detection and recognition as a function of the fraction of the target perimeter that can visually be resolved. It evaluates the target against its immediate background and does not analyze the information in the rest of the visual field. His analysis of the SEARCH_2 data yielded a distribution of fractional perimeter values similar to that resulting from a previous UK field trial. He argued that this finding supports the assumption that global scene analysis is not required to model observer performance for generic scenarios.

The second paper, "Image Structure Models of Texture and Contour Visibility", by W.S. Geisler, T. Thornton, D.P. Gallogly and J.S. Perry, of the University of Texas at Austin, USA, was presented by Bill Geisler. He argued that "bottom-up" mechanisms for grouping and segregation are absolutely essential to object detection and recognition. To recognize an object in a typical natural environment, the features of the object must be segregated from its local background. In most quantitative vision models the image is initially processed by channels selective along certain fundamental stimulus dimensions such as spatial frequency and orientation. These channels generally contain a nonlinearity, such as full-wave rectification, so that they signal the local contrast energy within the passband of the channel. Another stage of linear filtering, followed by a simple edge finding or thresholding mechanism, is then applied to the channel outputs to find the texture boundaries or regions. Although these channel-energy models have been successful in predicting texture segregation and discrimination performance for some classes of stimuli, there are large classes of stimuli that are readily segregated by human observers but which cannot be segregated by channel energy. Dr. Geisler demonstrated that a very simple image structure model, that combines local measures of proximity and continuation, can account for human ability to detect random contours that are representative, in complexity and uncertainty, of those occurring in the natural environment.

The third paper, "Comparing Human Target Detection with Multidimensional Matched Filtering Methods", by W.K. Krebs, D.A. Scribner, J.S. McCarley, J.S. Ogawa, M.J. Sinai, of the Naval Postgraduate School, Monterey, California, and of the Naval Research Laboratory, Washington D.C., USA, was presented by Kip Krebs. He compared the performance of a two-dimensional matched filter (spatial) optimized for a specific target and background power spectra, to the performance of human observers performing a standard search and detection task on cluttered multimodal images. False alarm and target detection probabilities were computed and results were plotted on a Receiver Operating Characteristic (ROC) curve. The results of the matched filter were similar to the results of the observers, indicating that the matched filter may be a good predictor of human performance. The matched filter approach has three obvious advantages. First, it provides a metric that can be used to evaluate different sensors. Second, it quantifies the effectiveness of image enhancement techniques, thus allowing for direct comparison of various enhancement algorithms. Third, it may have the ability to predict human visual performance across a variety of background and target conditions.

The fourth paper, "Detection of Low-contrast Moving Targets", by J.P. Mazz, R.W. Kistner and W.T. Pibil, of the U.S. Army Materiel Systems Analysis Activity, USA, was presented by John Mazz. He reported the results of an experiment that was designed to investigate the effects of target motion on the probability of target detection, for search with the unaided eye. The parameters of interest were background, target size (simulated range), target contrast, and velocity. Targets with near-equal contrast at identical range and angular velocity yielded widely different probabilities of detection. However, within a specific background region, contrast had a significant impact. Dr. Mazz argued that this localized impact of target contrast indicates that further improvements in search and target acquisition modeling require the evaluation of scene-content's impact on target detection (i.e., what about the scene leads an observer to the vicinity of the target.) For low-contrast targets, he observed that scene content has even greater impact on detection. This result agrees with the results of G. Aviram and S. R. Rotman (paper 7, Session I), who found that, for targets with low visual distinctness, the global clutter level determines the detection performance.

The fifth paper, "Validation and Verification of a Visual Model for Central and Peripheral Vision", by E. Peli and G.A. Geri, of the Harvard Medical School and Raytheon Training Inc., USA, was presented by Eli Peli. He reported the results of an evaluation study in which he compared the appearance

of an image viewed at various distances with simulations of that image corresponding to the same distances, generated with model a multi-scale vision model that applies a threshold (i.e. nonlinear) contrast sensitivity function (CSF) and a locally normalized, band-limited contrast definition. The model closely predicts observer performance, both for central and for peripheral viewing. This indicates that the use of a nonlinear CSF in combination with a locally normalized contrast metric may be valid. It also appears that the differences in image detail across wide-field images can be modeled using a single eccentricity-dependent parameter in addition to the foveal CSF.

The sixth paper, "Modelling of Target Acquisition within Combat Simulation and Wargames", was presented by the author, Jan Vink, of the TNO-FEL, The Netherlands. He addressed some of the limitations and problems of the current implementation of the target acquisition module in combat simulation and wargames.

The seventh paper, "The Deployment of Visual Attention: Two Surprises", was presented by the author Jeremy M. Wolfe of the Brigham and Women's Hospital and Harvard Medical School, USA. He presented the results of some visual experiments that demonstrate that covert attention is deployed at random among candidate targets, without regard to the prior history of the search. Only one object can be recognized at one time. Rejected distractors are not marked during a search, and may therefore be reinspected at a later moment. The author suggested that this surprising limit on our abilities may be based on a trade off speed for apparent efficiency.

SESSION III: SIMULATION ISSUES

The six papers in this session, that was chaired by Mark L.B. Rodgers of the Defence Clothing & Textiles Agency, UK, report studies that employ photosimulation and synthesized imagery in the design or assessment of CCD measures.

The first paper in this session, "Computational Models for Search and Discrimination: An integrated approach", by A. Copeland and M. Trivedi of the University of California at San Diego, USA, was presented by Mohan Trivedi. He conducted two different types of psychophysical experiments to generate quantitative measurements of perceived target distinctness for comparison to various computational target distinctness metrics. The first experiment involved paired comparisons of image stimuli that contain a target pattern embedded in a natural background pattern. For each pair of stimuli, the observer was required to select which of the pair possesses a target that is more distinct. By combining the decisions from a number of observers, he obtained numerical scale values for the relative levels of perceived target distinctness in the stimuli. These psychological scale values were compared to the computed values of different target distinctness metrics. The second experiment utilized image stimuli that contain several target patterns embedded in a background scene at random locations. The observer needed to perform both search and discrimination. The fixation point data from the observers were used to compute various statistics for each target indicating how easily the observers located it, including the likelihood the target was fixated or identified and the time required to do so. These computed statistics served as another quantitative basis for evaluating the relative effectiveness of target distinctness metrics at representing perceived target distinctness. For both experiments, he used the level of correlation with the psychophysical data as the basis for evaluating target distinctness metrics. Overall, of the set of target distinctness metrics considered, a metric based on a model of image texture was the most strongly correlated with the psychophysical data.

The second paper, "Depth Perception Applied to Search and Target Acquisition", by W.R. Watkins and L. Alaways of U.S. Army Research Laboratory and U.S. Military Academy, USA, was presented by Wendell Watkins. He performed a visual search experiment using both single and wide baseline stereo imagery. His image set contained the same scene both with and without camouflaged human targets present. He observed significantly longer search times for scenes where no target or a false target is detected. Second, he found little difference in total search time for one or many detected targets. Third, as the normalized time that a scene is viewed increased, the probability of false target detection also increased. He concluded that the use of stereo vision for reducing the clutter level in search and target acquisition tasks has promise, but requires care in assessing. He argued that it cannot be done for short range targets without using multiple fields of view.

The third paper, "Methods for Deriving Optimum Colours for Camouflage Patterns", by K.D. Mitchell and C.R. Staples of the Defence Clothing and Textiles Agency, UK, was presented by Kevin Mitchell. He argued that the design of scenario specific camouflage patterns involves two major factors: *(1)* the multi-level structure of a background and *(2)* the many colours present. The camouflage pattern should match the background at multiple levels of resolution, in order to be effective at various ranges. The design method needs to reduce the many hundreds of colours that usually occur in a scene to a workable number, usually between three and six. He presented a routine to derive optimum colours for a camouflage pattern using calibrated digital imagery. Optimised colours used in a pattern can reduce the ranges at which targets become visible in specific scenarios.

The fourth paper, "The Development of an Image Manipulation Facility for the Assessment of CCD", was presented by the author, Anthony W. Houlbrook of the Defence Clothing and Textiles Agency, UK. He discussed the benefits of photosimulation techniques as an alternative to performing live observer trials. The greater control over photosimulations allows an increased level of confidence in the results of any comparisons. It also requires less time in the field for a smaller number of personnel. The next step along this route is a method that requires no time in the field. Virtual reality systems, however, do not yet produce the level of realism required. An alternative, he started to develop a system to place targets generated by VR software into a scene recorded photographically. This system should digitize a slide of a background scene in a controlled manner and allow the realistic implantation of an artificially created target. Reproduction can then be achieved using a calibrated film printer. The majority of the reprinted scene should remain identical to the original slide. He discussed the methods used to enable the calibration of the equipment used, and the process of comparing information from digital rgb and Lab colour spaces.

The fifth paper, "A Physics Based Broadband Scene Simulation Tool for CCD Assessment", by I. R. Moorhead, M. A. Gilmore, D. Oxford, D. Filbee, C. Stroud, G Hutchings, and A. Kirk, of the Defence Evaluation and Research Agency, and Hunting Engineering Ltd, was presented by Ian Moorhead. He introduced the synthetic scene simulation system (CAMEO-SIM), that has been developed, as an extensible system, to provide imagery within the 0.4 - 14 micron spectral band with as high a physical fidelity as

possible. It consists of a scene design tool, an image generator, which incorporates both radiosity and ray-tracing processes, and an experimental trials tool. The scene design tool allows the user to develop a three-dimensional representation of the scenario of interest from a fixed viewpoint. Target(s) of interest can be placed anywhere within this 3-D representation and may be either static or moving. Different illumination conditions and effects of the atmosphere can be modelled together with directional reflectance effects. The user has complete control over the level of fidelity of the final image. The output from the rendering tool is a sequence of radiance maps which may be used by sensor models, or for experimental trials in which observers carry out target acquisition tasks. The system is intended as a tool for assessing the effectiveness of air vehicle camouflage schemes. However, the software is sufficiently flexible to allow it to be used in a broader range of applications, including full CCD assessment. He reviewed the verification tests that have been carried out, and described a validation programme based on *(1)* simple scenes, *(2)* neural nets to evaluate the higher order image statistics, and *(3)* comparison of observer performance in a real versus a simulated scenerio.

The sixth paper, "An Investigation into the Applicability of Computer-Synthesised Imagery for the Evaluation of Target Detectability", was presented by the author, Mark Ashforth of the Defence Clothing & Textiles Agency, UK. He found a marked difference in observer performance for the detection of targets in a real scene and the detection of targets in computer-generated (synthetic) images. Since synthetic imagery is increasingly used to assess CCD effectiveness, this is an important result. He explained this finding by the fact that there is less detailed clutter in synthetic images, which alleviates much of the decision-making an observer has to undergo in detecting a target in a real-scene image. In the synthetic case, the target is either seen or not seen, and there is much less uncertainty. This uncertainty, which attends real target detection, swamps any measurable influences on an observer's relative performance in the real-scene case. He concluded that computer-generated images used for the evaluation of low-contrast target detection should contain much more clutter detail than at present.

CONCLUDING REMARKS

Apart from visual conspicuity, which has been defined elsewhere[3], there appears to be no simple and efficient psychophysical method to quickly and reliably assess the effectiveness of CCD measures, both in the field and on (simulated) imagery.

A quantitative model of the human visual search and detection capability as-a-whole, that, given an arbitrary visual input to the eye, reliably predicts detection performance, is still a distant ambition. Factors that are known to be important in visual search, like complex pattern perception, visual attention, learning, and cognition, are still not addressed in most modeling efforts. Using a very simple image structure model, Dr Geisler (second paper of Session II) clearly showed the effectiveness of this approach. Furthemore, future research should address effective models to predict the rate of false alarm from image properties by deploying top-down knowledge.

The relative succes of local target distinctness measures as predictors of the human visual search and target acquisition performance may seem somewhat surprising, given the results of prior laboratory studies that indicate that the overall structure of a scene should determine this capability to a large degree[5]. The global structure (clutter) of the scene seems to influence detectability only for targets with low visual distinctness. Previous research has shown that observers tend to use a fixed amount of time for the initial inspection (one global scan) of a newly presented scene[2]. For a given field-of-view size, they will make large saccades and take long glimpses when the scene is relatively empty, and they will make small saccades at a fast rate when the scene contains a lot of detail, thus keeping the total time required for global inspection constant. If the target is sufficiently distinct there is an appreciable chance that it will be noticed when the fixation is in its vicinity (for less detailed scenes, fixations will be widely spaced, and the visual lobe of the target will be relatively large; for detailed/cluttered scenes the visual lobe of the target will be smaller but the fixations will also be closely spaced, thereby keeping the probability to detect the target in a single fixation near the target constant). If the single-glance detection probability is a function of the local target distinctness, and the time required for an overall scan of the scene is nearly constant, we may expect the aformentioned result, that local distinctness determines search time. If the target is of low visibility, intense scruntiny is required to find it, and this relation need no longer hold. Jeremy Wolfe (seventh paper in Session II) showed that obervers tend to reinspect parts of a scene they studied before. This behaviour may have ecological validity, since it obviously makes sense to monitor regions where predators (or nowadays cars) may appear, even if they were not there at initially. Hence, observers may decide to stop searching when one or more targets are found during an initial scan of the scene, and probably continue searching when no targets are found. Eventually, they may decide to pick the most likely of a number of candidate targets (image details that canot be identified with certainty, but that have some characteristic features of a target). Therefore, as Witus and Gerhart suggested (12th paper of Session I), the study of false alarms and eye movements may provide further insight into the processes underlying visual search and detection.

REFERENCES

1. Toet, A., Bijl, P., Kooi, F.L., and Valeton, J.M., "A high resolution image data set for testing search and detection models", TNO-HFRI Report TM-98-A020, 1998.
2. Toet, A., Bijl, P., Kooi, F.L., and Valeton, J.M., "A test of three visual search and detection models", TNO-HFRI Report TM-98-A038, Soesterberg, The Netherlands: TNO Human Factors Research Institute.
3. Toet, A., Kooi, F.L., Bijl, P., and Valeton, J.M., "Visual conspicuity determines human target acquisition performance", Optical Engineering, 37(7), 1969—1975, 1998.
4. Toet, A., "A comparative study of some target signature evaluation methods", TNO-HFRI Report TM-99-A053, 1999.
5. Bijl, P. and Valeton, J.M., "Observer experiments with Best Two thermal images. Part 2: Terrain interaction and target motion", TNO-IZF Report IZF-1992-A-34, 1992.

PERCEPTION STUDIES

Barbara L. O'Kane, Ph.D.
U.S. Army CECOM RDEC
Night Vision & Electronic Sensors Directorate
Ft. Belvor, VA USA 22060-5806
E-mail: okane@nvl.army.mil

David Bonzo
E-OIR Measurements, Inc.
Courthouse Road
Spotsylvania, VA USA
E-mail: dbonzo@eoir.com

James E. Hoffman
University of Delaware
Lincoln, DE
E-mail: hoffman@udel.edu

1. SUMMARY

Most United States ground-to-ground and air-to-ground military target acquisition is performed with thermal sights on today's battlefield[1]. It is extremely important, therefore, to correctly train, model, and understand the use of these systems to prevent fratricide and increase survivability. It is not possible to perform all the needed research by means of expensive field tests. Therefore, the use of perception studies has become popular for developing training, testing system designs, and assessing effectiveness of camouflage techniques. This paper discusses some challenges involved in perception studies that are conducted to gain insight into surveillance and target acquisition by military users of thermal imagery. The goal is to emulate as accurately as possible what a military observer will actually see and do when using a thermal sight. The issues discussed include prior training, panning effects on eye movements, and contrast and brightness controls. The latest advances in these areas and some remaining challenges are discussed.

Keywords: perception studies, target acquisition, thermal imagery, eye tracking, user interfaces

2. INTRODUCTION

Perception studies have been used for many years at the Night Vision & Electronic Sensors Directorate to provide data for target acquisition modeling. In combination with field tests, which can verify the overall results, perception studies can provide a controlled environment and flexibility to hybridize imagery, increase the number or types of subjects, and insert variables, such as simulated smoke, weather, or other image degradation[2]. The perception study methodology is not, however, without its detractors. There have been concerns that unless the operator is really in the field with his hands on the sensor that his experience is very different and that the results may not be comparable. Especially for camouflage evaluation, which is a specific example of perception studies, "photo-simulation," as it is often termed, has been called into question as a viable technique because of the need for highly similar circumstances in the laboratory and the field. We have addressed some of these problems with thermal imagery in the last decade and have had spin-off benefits that have been deployed throughout the world.

One of the first questions which comes to mind when preparing for a perception study is how and how much to familiarize the test subjects, also referred to as observers, with the camouflage system being tested. If they know, for example, that a certain camouflage net will cause the silhouette to be more rounded, should we allow learning during the perception study trials or show them straight off what the camouflage net looks like?

A second problem is that field studies are often performed by having the observers search a field of regard of say 30-120 degrees, composed naturally of many fields of view. But most perception studies are conducted on simply one field of view, known as "in-field-of-view search". Are these two scenarios related? Can the results of a field of view test be considered relevant for a field of regard search task such as would be encountered in the field?

When an observer is in the field and has his hands right on the controls of the sensor, he can adjust at any moment to bring out certain details, change the magnification, or switch polarity from black hot to white hot. Unless a perception test user interface can allow that kind of observer interaction, it is not actually simulating what would happen in the field. Would the camouflage net be as visible if the observer were given just a single, albeit allegedly "optimal", image?

A related issue is how much time to allow or to force an observer to search each image in the data set. Perception studies are performed with a wide variety of methodologies in terms of the time allotted. Each will have a dramatic effect on the results, and various methodologies may access a different stage in the attentional and perceptual processing of the scenes[3].

Each of these topics can be approached from the purely scientific standpoint or the military application view. The choices to employ a particular methodology can result in a study that provides tremendous scientific relevance but may or may not be useful to the designer of the camouflage or the builder of the sensor. On the other hand, as scientists we do not wish to perform scientifically mundane, inadequate, or uninteresting experiments; our desire is to contribute to the body of scientific understanding with each experiment that is performed. Thus, there is a balance between general scientific inquiry and the applications of perception studies to military

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

scenarios and the assurance that the customer is getting the biggest "bang for his buck."

3. OBSERVER TRAINING

There has been an operating perception laboratory at the Night Vision & Electronic Sensors Directorate for many decades. In the 1980's Dr. James Howe performed interesting studies on the effects of sensor parameters and target signature statistics on the probability of identification by military observers[4]. A typical study involved observers being shown images on the computer screen of various tactical vehicles, such as a T62 (a Soviet tank) or an M60 (a US tank) and being asked to identify them as quickly as possible. The images were generally taken from real thermal imagery that had been modified (degraded) to appear like sampled images or with diminished target signatures as shown in Figure 1. The experimenters noted almost immediately that the observers were not able to identify even the most close-up and undegraded images, because they had never been trained to identify thermal imagery. Their training had consisted of flash cards of line drawings reflecting the physical characteristics of the vehicle, photos showing the vehicles in daylight or slides taken with a camera attached to an image intensifier, a totally different technology than thermal imaging. The training of the military users of thermal systems had not included learning the special characteristics of each target as seen through a thermal imager.

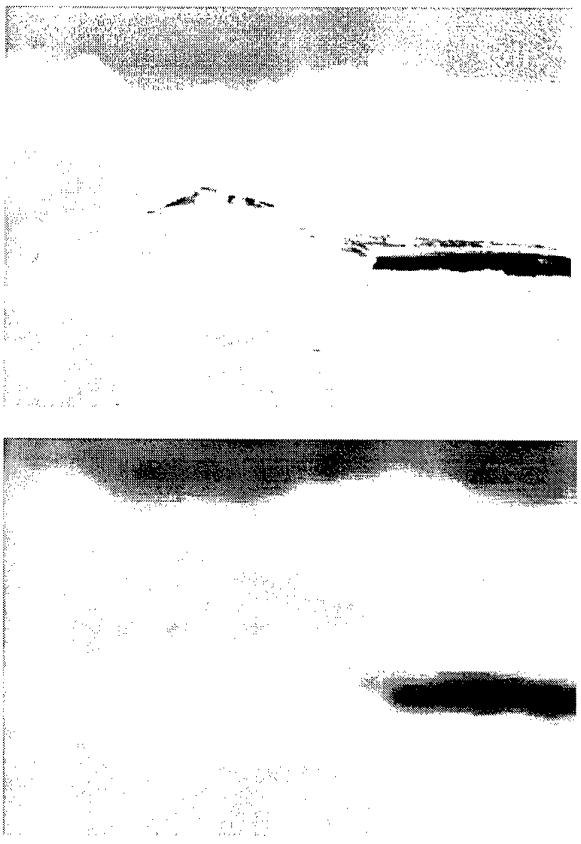


Figure 1. Picture of two images. Above, highly resolved and below, significantly degraded with simulated sensor effects.

Therefore, the experimenters were not able to determine the effects of sensor parameters or target signature qualities at all, due to the low training level. All these experiments would have failed had the experimenters not developed a special training program designed for the purpose. It was very simple, and fit on one of the old 5.25" floppy disks, which turned out to be very useful at the time. The training package simply showed each of the targets with the nomenclature displayed ("M60" or "BMP") and arrows pointing to the tracks, turrets and unique hot spots. Before participating as observers in the perception studies, the military observers were required to study the imagery and then take a test to ensure a 90% or greater training level criterion had been achieved. This training improved the potential of the perception studies greatly.

But what happened after that was quite surprising. The experimenters were asked by the military participants if they could please take a copy of the training package back with them to their units because it was the "best" thermal training they had ever had in target signatures. The soldiers would then return to their posts and tell others about the training. These soldiers would then write or call us requesting the training software and before long we had sent out hundreds of these packages.

However, the issue of how to identify targets in thermal imagery remained a subject for scientific research. Through perception studies over about four years it was determined that the best approach to understanding thermal signature identification and training was in terms of Recognition-by-Components theory[5], developed by Prof. Irving Biederman of the University of Southern California, Berkeley. In 1996 the development of the training program commenced and it is now in distribution to the US military.

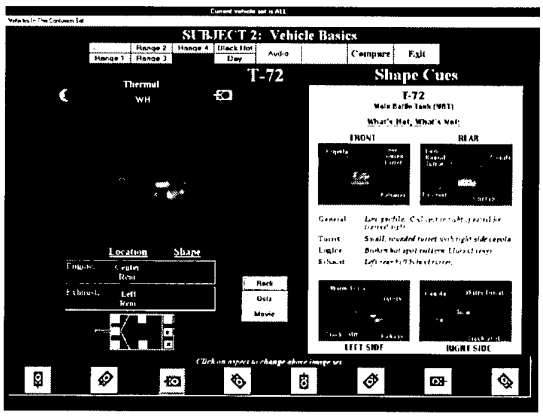


Figure 2. New training package vehicle basics module.

During the development of the program, a First Article Test of the training was performed at Ft. Hood with 109 soldiers to determine how significant an impact on combat vehicle recognition this program would have in the real military community. The observers took a pre-test before training on the 16 combat vehicles they were going to be trained on, which were the various vehicles expected in any present-day conflict involving former Soviet Union members or the U.S. The average probability of identification for the 109 soldiers was less than 20% correct. After four hours of training, the average target identification was almost 60%. Since that time we have trained many observers to 95% with one or two days of self-training. Since the distribution of the program to many bases of the US military in the last year, a pretest in the present environment would likely yield a higher score. Thus, in this way, a complete circle has been made, the perception tests are now affecting the performance of the troops through

the spin-off benefit of training and the training will undoubtedly change the results of perception tests.

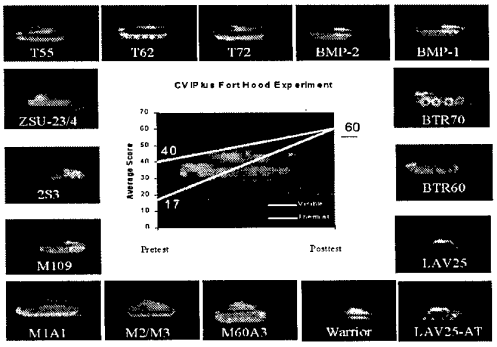


Figure 3. Results from large test before and after four hours of training.

What is significant about this perception study methodology and modeling is that observer training, or its inadequacy, can overwhelm the effects of the factors under study, such as sensor parameters (sampling, optics, blur, noise, etc.). We are presently working on similar training packages addressing the issue of search and detection. When this is done, it will not only change the results of perception studies and how they are performed, but may significantly change the way that observers acquire targets on the battlefield.

Less variability between subjects is found for the detection of targets than for the identification of targets. There is a much more inherently natural process involved in the discrimination of target from background than in the understanding and remembering of the cues and features involved in target identification. Nevertheless, with camouflage techniques, there will be greater difficulty with detection. This difficulty would be expected to cause greater variance among observers, since those with greater skill and experience with the particular camouflage systems and sensors may significantly surpass less experienced observers.

A related issue is how much familiarization an observer in perception studies should be given prior to his participation. To answer this question, it is useful to know what the customer perceives as the most significant question. Does the customer want to know how the camouflage will work the very first time it is deployed or how it will work after it has been deployed long enough for the threat to have familiarity with the item? In general, it is probably more realistic to assess camouflage assuming a basic familiarity with the item. If there is no familiarization with the system, then under the conditions of multiple trials normally conducted, the observer will learn during the testing what the item looks like and this will cause a systematic error in the data with trial number. Ordinary procedure, therefore, would cause us to familiarize observers with the camouflage techniques prior to testing.

A very famous report by the Harvard Business School published in 1939 reviewed the results of a 12-year study on the effect of attention and neglect on workers at the Hawthorne plant of the Western Electric Co., manufacturing equipment for AT&T[6]. The research seemed to show that performance improved with any kind of noticeable attention to or change in the subjects' conditions. Raising the illumination or lowering the illumination, reducing the work week or returning to the standard work week, introducing rest pauses or decreasing them, all caused an increase in productivity. When we perform our experiments and training, therefore, we are likely to improve performance over the usual for a soldier in the field. Are the studies then not valid for predicting real life operational performance? This is a reasonable question and one that we must seriously consider when reporting the results of our studies. It would seem to be true that our studies represent a "best-case scenario (or "worst-case scenario" when studying camouflage). The subject knows he is being carefully watched, he is glorified somewhat by this attention, and he is directly aware of the purpose of the experiment. Can this so-called "Hawthorne effect" be mitigated in any way? It is hard to imagine an experimental situation in which the subject does not feel that he is being given attention, that he is unaware of the purpose of the experiment, and that he does not feel glorified and in a different type of situation than usual, all causes for the Hawthorne effect's improvements in performance. The best way to approach this problem is to rely little on the absolute performance values, and more on the relative values between treatments of the camouflage or baseline item. Making statements about exact ranges and probabilities of detection is somewhat risky because perception studies do not represent the treatment of a soldier in the field.

4. FIELD OF REGARD SEARCH

Typically, perception studies have been conducted with single fields of view of various sizes. However, it is unusual for a battlefield or other ordinary field situation to involve a single field of view. Rather, a user of thermal sights will have a sector designated for search, which might be between 30 to 120 degrees. It is often found that field results for search and target acquisition are lower than the model predictions. If we continue to study only field-of-view search, we may not be emulating the most important of situations, and may be drawing conclusions that are not relevant to the battlefield case.

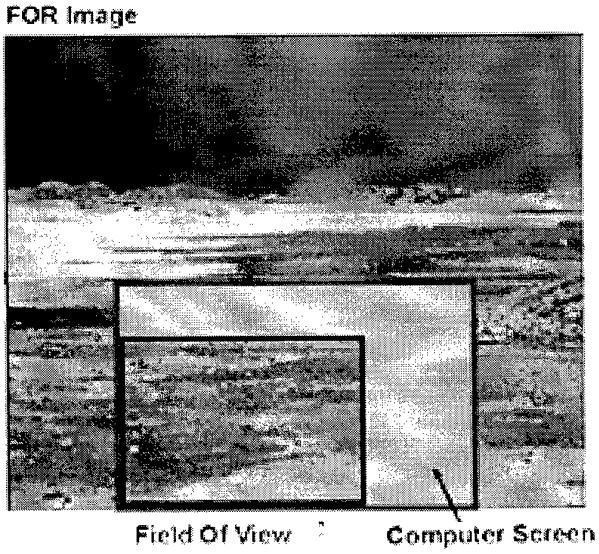


Figure 4. A large field-of-regard image being moved horizontally left while viewed through a smaller field-of-view.

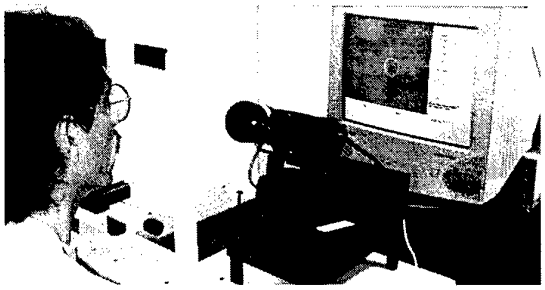


Figure 5. Setup for eye tracking experiments.

One experiment in particular raised questions concerning the cause of relatively low detections with thermal sights. A concern that naturally arose was whether the panning itself or its speed was causing reduced detections. One way to address this, which we pursued, was eye movement tracking during field of regard search. One of the authors, Dr. James Hoffman, professor at the University of Delaware, conducted an experiment with NVESD and demonstrated that panning field of regard search (Figure 4 and 5) imposed specific eye movement patterns which involved an up and down movement near the leading edge, as shown in Figure 6. In contrast, a step-stare mode, in which one scene after the next was displayed sequentially, with the same duration per field of view as in the panning mode, demonstrated a totally different eye movement pattern as shown in Figure 7. In the step-stare approach the eye movements tended to be concentrated around the center of each field of view with decreasing cumulative fixations at greater x and y eccentricities[7].

A moving scene in front of the eye prevents the normal modes of eye movements, which involve peripheral detections and saccades to various parts of the same scene. In fact, the scene is different by the time the next eye movement occurs in a panning field of regard search. What passed quickly across the fovea may be in the periphery and as the fovea moves to fixate on a target in the moving scene new information is coming into the periphery still. Clearly, this is not the case for a field-of-view search. In this case, the eye movements range out from the center and the fovea is not being required to parse a scene moving in front of it. Peripheral vision can be used in its normal mode, although the whole field of regard scene is not available at one time.

On a more general level, the brain is able to comprehend a scene at a glance[8] when in a normal mode and saccade the eyes to the most promising target area. With a moving window on the world, there is difficulty with establishing a "world" with placeholders (as Dr. Biederman has called it), which can be foveated and more highly resolved later. Instead, the field of view is moved in a sometimes more, sometimes less, random fashion throughout the scene and the benefits of peripheral vision only extend to a strip on the other side of that in which new information is entering. When evaluating camouflaged items, this type of searching will result in a poor detection rate relative to field-of-view search and introduce a great deal of noise into the equation[9].

The question persists, does the experimenter opt for more realism in the testing environment (field-of-regard search) or a more conservative test of the camouflage system being evaluated (field-of-view search)? Some experimenters have suggested the latter, but there may be two different questions being asked and answered by the different methodologies.

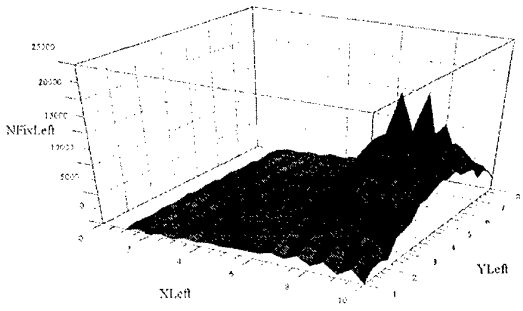


Figure 6a. New scene information is entering from the right.

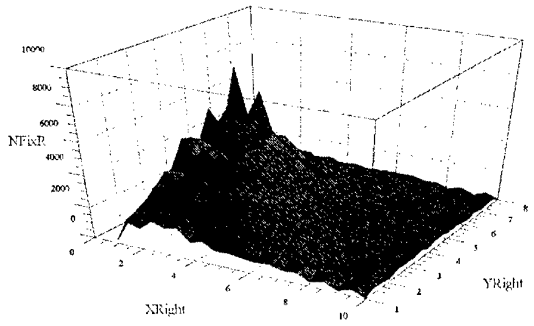


Figure 6b. New scene information is entering from the left.

Figure 6. Cumulative fixations in slow panning mode for scenarios with new information entering from the right and left. The eye is most often positioned near the center of a line about 1 inch from the leading edge at which the new information is entering the screen.

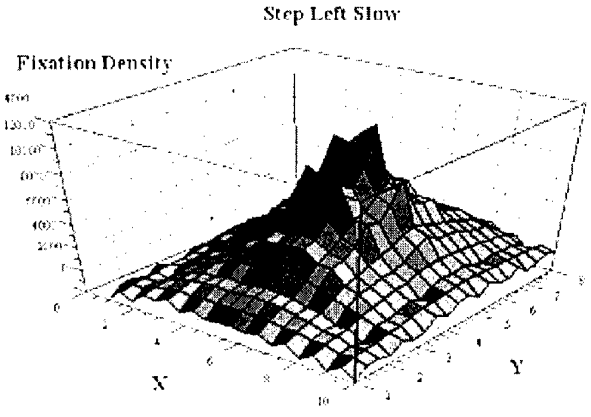


Figure 7. Cumulative fixations for step-stare mode of searching a field of regard. The eye is most often positioned near the center of the field of view with excursions outward.

The field-of-view search methodology can answer the question: "If there is enough situational awareness or some other cue which allows the sensor to be directed at the camouflage, will it be recognized as an item of military significance?" This question is very conservative for camouflage technology evaluation, but will give the clearest answer. The detection of camouflage is clearly a function of the eccentricity of the target from direct fixation. Biederman et al.[7] showed that large targets (i.e., those subtending at least 3 degrees in the visual field) lying within 5 degrees of the fixation point were readily detected 80% of the time with only a 150ms exposure. However, small, camouflaged, or unexpected targets required a much closer distance to the point of fixation to readily detect the target. This phenomenology is also the basis of the methodology for judging the effectiveness of camouflage developed by Toet et al.[10]

The field-of-regard methodology will answer the question: "In a fairly realistic environment, how likely is this target to be found and discriminated from the background relative to a non-camouflaged target, given that the observer may not actually have the target in the field of view for very long?" This will add considerable statistical noise to the data because there will be many more non-detections than would occur with the field-of-view methodology. These non-detections will be due to the extra time spent looking at the parts of the scene that had no target and to the small amount of search time actually spent with the target in the field of view. While perhaps more realistic relative to a battlefield in one sense, in another sense there would be many more sensors out there looking, adequate situational awareness and other means of determining the possible location of the target than panning one sensor through a large field of regard.

King, Stanley and Burrows[11] have an excellent discussion of the issues involved in field of regard search experiments for use in evaluating camouflage. We have found, as have others[8,9,11], that the search time required for fields of view in which observers do not detect the target is much higher than are search times for those scenes in which a target is detected. Therefore, all of the empty scenes in fields of regard will take a significant amount of the attentional capacity of the observer and greater reaction times will occur. These are not necessarily a reflection of the camouflage method, but is the "confounding effect of a large, complex field" on the evaluation of camouflage[9]. We have found that observers need a minimum amount of time with their eyes on the target for a detection to occur, not just the eyes passing over the target, as Nodine[12] also found. Therefore, care must be exercised in conducting perception studies with fields of regard to evaluate camouflage.

Studies are currently being designed in our laboratory to investigate the difference between field of regard and field of view search and to apply the appropriate factors to the results. One method that is being planned is to present the field of regard in two ways, one with the scene completely on the screen, and the other with the panning mode. This comparison and that of changing the size of the field of view that encompasses the target can attain a modeling of the field of regard search times and probabilities relative to field of view search with different size fields of view. It is hoped that this understanding will significantly improve the modeling of search and target acquisition for fields of regard. Current search models predict search time for field of regard as a multiple of the search time for individual fields of view.

We know however, that search is a very different process in the field of view and field of regard modes (See Figures 6 and 7.) Therefore, it is likely to be inappropriate to use a multiplication model to leap from field of view to field of regard modeling of search times. In the comparison of perception studies and field trials, which normally do include more than one field of view, there continues to be an issue. If we do the most efficient perception study, it will be a field-of-view test. However, developers of camouflage will typically, inevitably, and rightly assert that their camouflage would be more effective in a field of regard search. Certainly it will be detected less often[11] but so will the baseline. The question is whether the field of regard clutter more effectively obscures the camouflaged item than the baseline item. And if so, is that as practically important as whether the item is difficult to see when the observer is looking directly at it? A serious developer or user of camouflage would want the item to be difficult to see in the event that it is looked at directly since the user cannot be assured a cooperative background during use. This need would mandate a field-of-view perception test to estimate the stealthiness of a particular camouflage item. However, a field of regard perception study comparing the detectability of the baseline and camouflaged items may be of interest to developers in deciding which technology to pursue.

5. GAIN AND LEVEL CONTROLS

When experiments are performed using imagery collected in the field and brought back into the lab for further testing, the format and display of the imagery is a great concern. With 1st Generation FLIR in the tactical version there is no output imagery from the FLIR before the display. Some laboratory versions have an RS-170 output port, but this is very rare. In some other cases, an E-O Mux has been created and used for testing but this is generally only a fair representation of what the observer actually would see through the sight in the field.

The advanced "2nd Generation FLIR" is designed with Automatic Target Recognition systems in mind and therefore has a 12-bit digital output port. When collecting field imagery it is possible to obtain this 12-bit output and then to show it on a CRT in eight (8) bits. As is well known, there are many ways to show the 12 bits on a CRT. In the field, the observer has two choices, automatic or manual gain and level controls.

The system designers have developed an algorithm that is used in the automatic gain control that processes the 12 bit data to optimize the appearance of the image. What one sees in the FLIR display tends to be the best image of the background, the leaves on the trees and all the cues needed for navigation through the terrain. The other mode of operation is the manual gain and level control, which can be used to reduce the overall brightness of the scene and enhance the contrast between various points in the scene, with certain limits and restrictions put into the proprietary schemes.

It had long been suspected that the automatic gain control was not optimal for detection but personal computer speed for processing imagery did not permit a smooth emulation of manual gain and level controls for widespread use in perception studies. While certain very expensive workstations were capable of producing a somewhat realistic rendition of sensor controls, perception studies were cumbersome to conduct due to the fact that only one such machine was available and there was a very high maintenance requirement. With the latest upgrades in PC speed and processing power, we are now able to do a manual gain control on an image in real-time field of view search. This has enabled us to determine that the manual gain control does indeed improve performance, sometimes markedly over the automatic gain control algorithm presently installed in some advanced sensors. Figure 9 shows the results from a recent study demonstrating the better performance when allowing the observer to adjust the gain and level controls himself over

using the automatic gain control in a field of view search of high clutter scenes.

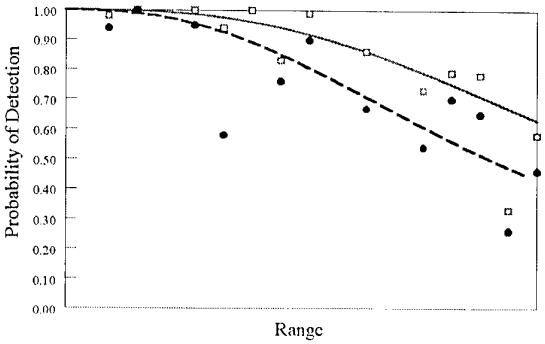


Figure 9. Differences in probability of detection as a function of range between automatic and manual gain control during a perception test on detection of a tank.

The automatic gain control is well-suited to maintaining stability of the scene dynamics on the display when the platform is moving through terrain. While on the move, the operator cannot make manual adjustments to the display to optimize each picture as the thermal scene varies from location to location, through sky, horizon, forest, tree line, road, and grasslands. The automatic gain control adjusts the scene to provide a consistently clear picture of the terrain and potential targets. The manual gain control can provide the detail and the ability to discriminate less obvious targets from the background in a cluttered scene (Figure 9.) From the standpoint of a conservative evaluation of camouflage, the manual gain control would add a more stringent criterion than would a pure automatic gain control evaluation.

6. TIME RESTRICTIONS

Another issue to be resolved by the experimenter before running a perception study to evaluate camouflage is whether to allow the observer unlimited time or to limit the time allotted. Discussions of each of the potential general methodologies follow:

6.1. Glimpse

This methodology is intended to prevent eye movements and to reveal automatic detection processes in the visual system. The observer would normally fixate a central point, be presented with the stimulus for less than 500 milliseconds, and then be presented with a mask. The observer would respond by indicating whether there was or was not a target in the field of view, or at a specific location in the field of view.

The disadvantages of this methodology stem from the fact that it does not represent a real-world type scenario. Except for the case of fast-moving jets, few scenarios exist in the military involving a single glimpse. The jet scenario differs from the glimpse methodology because the platform is also moving relative to the scene. Glimpse methodology taps vision mechanisms involved at the first stages of processing and in active visual memory (the visual memory sketchpad)[3], and has been very useful for studies of scene understanding, semantic and syntactic processing of scenes, and camouflaged versus baseline targets[8,11].

Whereas the glimpse methodology has basic scientific appeal in the automatic detection and rapid visual processing of information, the drawback of having relatively low face validity relative to military tasks makes it a less optimal choice for perception studies in most cases.

6.2. Time Restrictions versus Unlimited Time

A methodology which restricts time is intended to emulate a typical battlefield scenario in which the observer has a limited amount of time to detect a target, somewhere between one and 30 seconds. In field manuals, there is a time requirement for the weapon system operator to make a detection and fire the weapon and this time limit can be applied to the perception tests and provide the user, the developer, and the scientist with a basis for assessing the camouflage relative to the baseline untreated system. In most of the perception studies conducted at Night Vision in the recent past, we have found that ten seconds per field of view is generally adequate for most scenarios. It has been anecdotally observed frequently that ten seconds is usually the limit for correct identification as well. After ten seconds, the observers generally will be incorrect in their identification of the target, and few correct detections are made after ten seconds in field of view search with unlimited time.

Adding a time limit such as ten or twenty seconds may reduce the false alarm rate and cut off search of fields of view which have no target, but will rarely influence the detection of baseline targets. In the case of effective camouflage, the time period may have to be extended to 30 seconds or beyond.

6.3. Forced Time

Another possible scenario to use in assessing camouflage is to force the observer to view an image for a certain length of time, for example, 60 seconds. What is likely to occur in this case is that the observers will not feel a sense of urgency and detection times may be much greater than would normally be the case, especially if notice is given of five or ten seconds left, as the end nears. Detection times would appear to be more accurate when observers are encouraged to work as quickly as possible and the reward of leaving the perception study sooner is coupled with detection performance.

7. OBSERVER MOTIVATION

It is important to motivate observers for their task. Normally, military observers are highly competitive with each other for scores. They want to know how they did relative to others as soon as they finish the test. It is normal practice in our lab to inform observers at the beginning of their participation that there will be a gift award given to the best performer. The score used to determine who will receive the award is calculated by dividing the number of correct responses by the average response time. Subjects are thus motivated to be quick as well as accurate.

8. SUMMARY

The approach taken in perception studies has been to perform research that can be directly related to a military operational scenario. When questions arise as to which of the potential approaches to use, the inquiry centers on how a particular perception study methodology would relate to a military operator's experience or to field manuals. By using this approach, the choice of methodology is based upon a principle that is not strictly academic, thereby reducing conclusions of purely scientific interest, but rather makes the link between the

laboratory perception study and the battlefield experience most robust.

Five general issues and recommendations for perception study methodologies were discussed: training, field-of-regard versus field-of-view search, gain and level controls, time limitations, and subject motivation. As technology allows greater capability to provide optimal emulation of military operational procedures, the perception studies can become more and more realistic. Recent computer speed upgrades have allowed advanced training, manual gain and level controls, and field of regard search with eye tracker. All of these enhancements contribute to perception experiments with greater face validity, which can be important for developers and users of camouflage in evaluation of systems.

9. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the data analysis and curve-fitting support of Dr. David Wilson. We also acknowledge the invaluable expertise of Mr. Dave Bennett of E-OIR Measurements, Inc. in the area of user interface development. Ms. Michelle Tomkinson assisted in the conducting of the eye-tracking studies. PM FLIR support to training is gratefully acknowledged. John O'Connor's technical consultation and excellent support to training and perception studies has been a great contribution.

10. REFERENCES

1. Herdman, R. "Who goes there: Friend or foe?" Technical Report No. OTA-ISC-537. US Government Printing Office, Washington, D.C. 1993.
2. O'Kane, B. L. "Validation of models through perception studies", in *Vision models for target detection and recognition*, E. Peli, Ed., pp. 192-218, World Scientific, Singapore, 1995.
3. Hoffman, J. "Stages of processing in visual search and attention", in *Stratification in Cognition and Consciousness*, B. Challis & B. Velichovsky, Ed., John Benjamins, Amsterdam/Philadelphia, in press.
4. J. D. Howe et al., "Thermal model improvement through perception testing", *Proceedings of IRIS Specialty group on Passive Sensors*, Infrared Information Analysis Center, ERIM, Ann Arbor, MI, 1989.
5. Biederman, I. "Recognition-by-components theory", *Psychological Review*, 94, pp 115-147, 1987.
6. Roethlisberger, F.J. and Dickson, W. J. *Management and the worker*. Cambridge: Harvard University Press, 1939. (Cited in Mouly, G.J. *The Science of Educational Research*, American Book Copy, New York, 1963, pp.443-446.)
7. Hoffman, J.E., O'Kane, B.L., and Tomkinson, M. "Eye movements during search of a large thermal field of regard", *Proceedings of the IRIS Symposium on Passive Sensors*, Vol. I, 1998, available at http://hoffman.psych.udel.edu.
8. Biederman, I. "On the information extracted from a glance at a scene", *Journal of Experimental Psychology*, 103, 597-600, 1974.
9. King, M.G., Stanley, V.G., and Burrows, G.D. "Visual search processes in camouflage detection", *Human Factors*, 26(2), pp. 223-234, 1984.
10. Toet, A., Kool, F.L., Bijl, P., and Valeton, J.M. "Visual conspicuity determines human target acquisition performance." *Optical Engineering*, 37(7), pp. 1969-1975, 1998.
11. Biederman, I., Mezzanotte, R. J., Rabinowitz, J.C.,, Francolini, C.M., Plude, D. "Detecting the unexpected in photointerpretation", *Human Factors*, 23(2), 1153-164, 1981.
12. Nodine, C.F., Carmody, D.P. and Kundel, H.L., "Searching for Nina", *Eye Movements and the Higher Psychological Functions*, J.W. Senders, D.F. Fisher, and R.A. Monty, Eds. Erlbaum, Hillsdale, NJ, 1978.

Lessons Learned in Developing and Validating Models of Visual Search and Target Acquisition

[1]Theodore J. Doll and [2]Richard Home

[1] Electro-Optics, Environment, and Materials Laboratory
Georgia Tech Research Institute
Georgia Institute of Technology
Atlanta, GA 30332-0841, USA
Phone: (+) 404 894 0022
Fax: (+) 404 894 6199 / 6285
E-mail: ted.doll@gtri.gatech.edu

[2] Defence and Evaluation Research Agency
Malvern, Worcester, WR14 3PS, UK
Phone: (+) 44 1684 896 933
Fax: (+) 44 1684 896 714
E-mail: rhome@dera.gov.uk

1. SUMMARY

Some shortcomings of past and current approaches for modeling human visual search and target acquisition (STA) are discussed. The effects of complex pattern perception, visual attention, learning, and cognition on STA performance are particularly emphasized. The importance of these processes is explained and approaches are suggested for modeling them. Guidelines are also provided for testing and validating models of visual search and target acquisition. These guidelines take into account the roles of pattern perception, visual attention, learning, and cognition in STA performance. The present paper also presents and compares alternative approaches to field testing for the purpose of model validation.

Keywords: search, target acquisition, perception, attention, learning, cognition, validation

2. INTRODUCTION

The military spends millions of dollars annually to build large-scale, system-level simulations of weapons and related systems. These simulations enable their users to understand how the systems will perform under conditions that would be impossible or extremely costly to produce in the real world. However, very little money is spent on system-level simulation of the one system that is key to all military operations – the human visual system.

System-level simulations of human vision could be useful in setting performance standards for both the naked eye and all types of sensors and systems in which the final judgement or interpretation is made by a human observer. System-level simulations of human vision would also lead to more accurate design requirements for sensors and camouflage, concealment, and deception (CCD) systems. A better understanding of the human visual system would also provide insights into how best to test and validate models of search and target acquisition (STA) performance.

Until recently, attempts to build general models of human observer target acquisition performance have met with only limited success. By the term "general", we mean models that accurately predict the detectability of (at least) military targets as viewed through a wide variety of sensors in a wide variety of backgrounds, without the need for calibration in each new situation. The difficulty no doubt stems in part from the inherent complexity of human perception and performance – but also in part from the manner in which the problem has been approached by the military R&D community.

Military-sponsored STA modeling has traditionally followed either of two approaches: (1) physics-based, or (2) simple models of human visual performance that emphasize only a part of the neural "machinery" involved in human STA. The physics-based approach is based on the idea that simply matching the target signature to the background clutter will suffice to deny detection. In spite of decades of research, this approach has failed. The reason is that no one has been able to determine to which aspects of the background clutter it is necessary to match to the target. It has proven impossible to match targets to all aspects of background clutter because clutter characteristics change over and within scenes (i.e., clutter is non-stationary).

Modeling efforts following the second approach – modeling only a limited part of the visual system – have typically emphasized the basic sensitivity of the eye to light, or at best, the basic spatio-temporal contrast sensitivity of the visual system. They typically pay scant attention to the important roles of complex pattern perception, visual attention, learning, and cognition in STA performance. Thus, they model only a limited part of the visual system. This state of affairs has occurred, in large part, because there has not been wide-spread understanding of the attentive, perceptual, and cognitive aspects of visual

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

performance and the role of learning in the military R&D community.

There is, however, a widening awakening to the role of attention, perception, cognition, and learning. Some of the papers in this conference attest to that fact. In his abstract, Al Ahumada remarks that "learning and memory components are required for a model that can accurately predict human detection in unpredictable backgrounds." In discussing shifts of attention during search, John Findlay suggests that "eye movements are programmed on the basis of a spatial salience map with both excitatory and inhibitory influences reaching it from feature maps", and "flexibility in search is provided through learning mechanisms." Commenting on the role of perceptual organization in contour and texture segregation, Wilson Geisler notes that "evidence suggests that more sophisticated models incorporating perceptual organization mechanisms will be required to predict human texture and contour segregation performance."

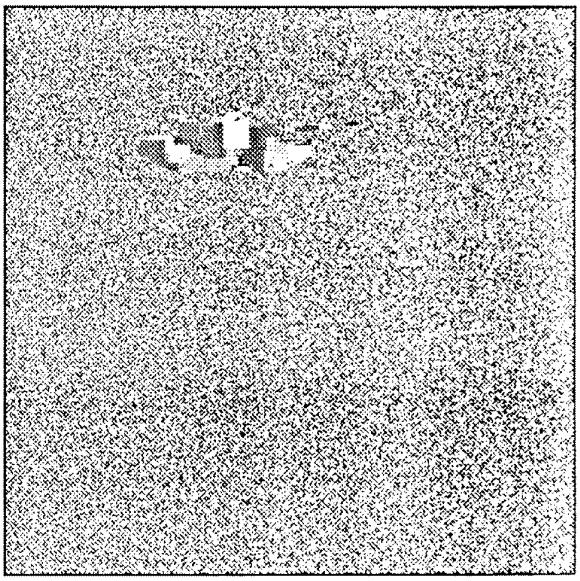


Figure 1. Tank and background with identical first-order statistics.

The cost of ignoring attention, perception, cognition, and learning is that the models developed have limited scope, and must be empirically calibrated for each new sensor technology, background environment, CCD technique, and level of observer experience. In the remainder of this paper we will explain which aspects of attention, perception, cognition, and learning we believe are most important and why they must be modeled in order to predict STA performance accurately and generally. We will also describe the manner in which these processes have been implemented in one model of human search and detection performance – the Georgia Tech Vision (GTV) model. [1]

Unfortunately, including all the relevant visual processes leads to very complex models that are difficult to validate, as Richard Hecker notes in his abstract. However, we disagree with Dr. Hecker's implication that higher perceptual processes like recognition, identification, and search can be eliminated from a model and still have it generate accurate predictions. We will therefore also discuss requirements for model testing and validation that take into account these higher level processes.

3. ROLES OF ATTENTION, PERCEPTION, COGNITION, AND LEARNING IN STA PERFORMANCE

3.1. Perceptual Organization

Walker and McManamey[2] point out that first-order statistics do not provide information about the spatial structure of an image. First-order metrics include the mean and standard deviation, as well as some less well-known metrics like the Doyle metric and measures of histogram similarity. The tank and the background shown in Figure 1 have identical means and standard deviations, and they're also identical in terms of the Doyle metric and histogram similarity. But they differ in terms of the arrangement, or spatial structure, of the pixels of various gray-scale values. The fact that the tank is clearly detectable from the background demonstrates that first-order statistics are not sufficient.

To account for the detectability of this target we must consider second-order metrics. The gray level co-occurrence matrix (GLCM) is one second-order metric; others include the correlation length and the co-occurrence matrix[3]. Both of these quantify the correlation between gray-scale values various numbers of pixel locations apart. Although the GLCM, correlation length, and co-occurrence matrix capture some of the properties that contribute to detection, they don't capture all of them. There are texture differences that humans can distinguish, but to which GLCM and correlation length metrics are insensitive.

The image in Figure 2a contains a texture irregularity that human observers can detect (note center bar-shaped region in the center of the image). However, most metrics and models of vision cannot detect this irregularity[4]. This is true of both single-stage, oriented linear-filter models and metrics like the correlation length, co-occurrence matrix, and GLCM. The reason for this is that the entire pattern is made up of the same texture elements – lines of the same length at different orientations. In addition, the probabilities of gray-level transitions from point to point are virtually the same in the center "irregularity" and the surround regions of the image. What distinguishes the center region is not the texture elements themselves, but their relationship to one-another. Note around the center region, that there are abrupt transitions in the relative orientations of the line elements. In the background, by contrast, the orientations of adjacent line elements change only slowly.

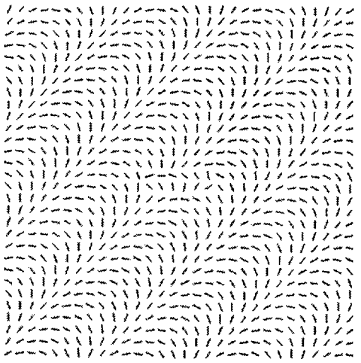


Figure 2a. Input image with texture transition near center.

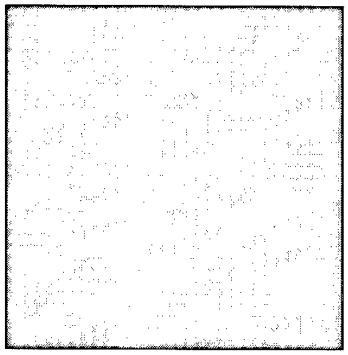


Figure 2b. Output of a single-stage, simple cortical cell, filter model.

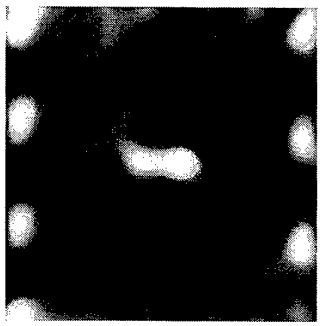


Figure 2c. Output of two-stage, complex cortical cell, vision model.

Another way of thinking about this pattern is that the center region is defined by a texture transition. In order to detect these subtle texture transitions, a vision model must have a second filtering stage, which models the outputs of complex cortical cells. Figure 2b shows the output produced by a model with only simple cortical cell (single-stage) filters, for the input in Figure 2a. Note that there is no differential signal that distinguishes the center irregularity.

The GTV model has a second filtering stage, as shown in Figure 3. Each first-stage output is routed to multiple second-stage, spatial-frequency band-pass filters. Depending on the version of GTV run, the second-stage filters may also be orientation-selective. The second stage filters smooth the outputs of each first-stage over regions of various sizes and orientations. This smoothing serves to identify the extent, or boundaries, of each type of texture identified by the first-stage filters. By comparing these boundaries, GTV can identify texture boundaries, as shown in Figure 2c.

But is the detection of such subtle texture transitions relevant to real-world CCD problems? Figure 4a shows a texture transition that might occur with a perfectly camouflaged vehicle positioned against a background of vegetation. When the vehicle is repositioned, there will be a phase mismatch between the texture of the vegetation and the camouflage pattern on the vehicle. Figure 4b shows a GTV output for this pattern, after the model is trained to detect similar phase-mismatched targets.

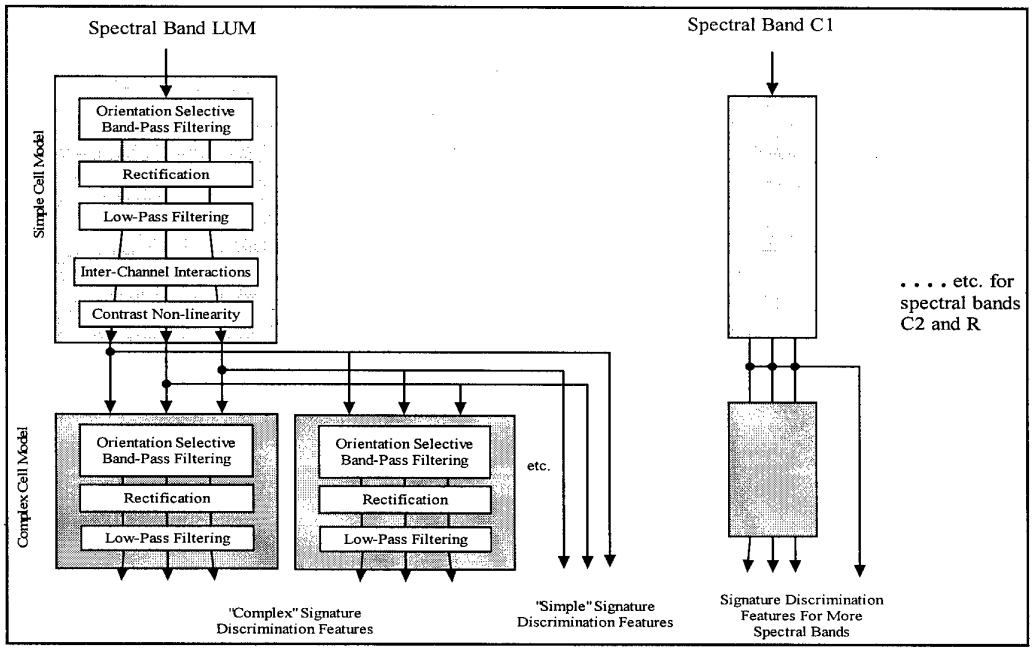


Figure 3. Schematic of GTV two-stage filter process, simulating complex cortical cell outputs.

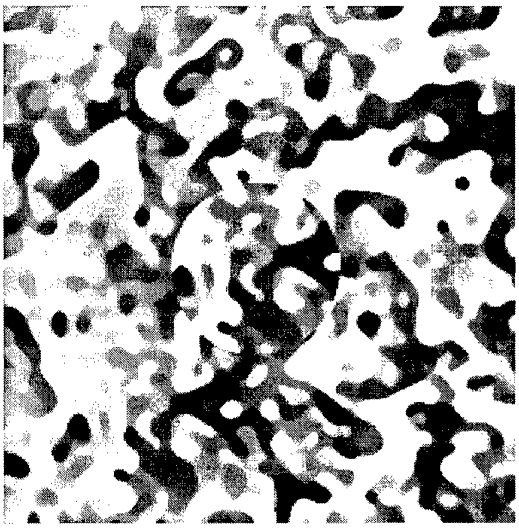


Figure 4a. Object perfectly matched in pattern and chromaticity to background.

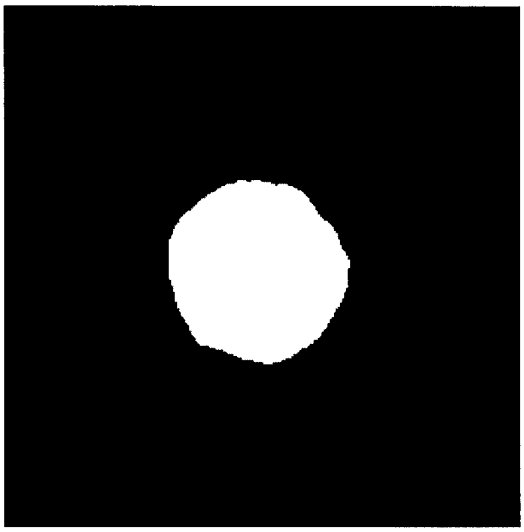


Figure 4b. Output of GTV identifying pattern in Figure 4a.

3.2. Attention and Search

There is substantial evidence that eye movements (saccades) during visual search are guided by preattentive (unconscious) processing of pattern information in peripheral vision. For example, recordings of eye movements over structured scenes reveal that the eye fixates on features such as edges and corners that are more likely to convey information than are plain surfaces.[5] In reading, the eyes of proficient readers search out larger words, which convey a higher degree of meaning than do small words, such as articles[6] .Visual search proficiency has even been used as a measure of peripheral visual acuity.[7–9]

The implications of this are:

- That clutter (i.e., the input scene) drives visual search.
- That successful search is a prerequisite for detection.
- The eyes fall on those objects that are most conspicuous.
- The assumption, often made in vision models, that search is random is false.

The first line of self-protection is not to be noticed in the first place, that is, to deny visual search. It's generally easier to prevent an observer from locating a target than it is to deny detection once he's looking directly at the target. This is especially true in medium- to high-clutter environments.

Explicit modeling of the effect of clutter on visual search is therefore necessary to accurately predict target acquisition. High clutter in an image reduces probability of locating the target, given limited search time. The GTV model predicts the fixation locations based on the spatial and temporal contrast of objects in the input image. A salience map is generated by using multiple-channel, quasi-linear filtering mechanisms. This map also serves as a basis for segregating the input scene into areas of interest for further (attentive) processing.

Another aspect of search that affects target acquisition is the temporal sequence of eye fixations in a scene. A wealth of data shows that human observers tend not to immediately re-fixate on objects when inspecting a scene.[10–12] The GTV model includes a systematic search routine which simulates the fact that observers tend to disregard objects that they have recently fixated and determined not to be targets. Thus, if an object has a high probability of fixation on one glimpse, and it is determined not to be a target, it will be less likely to be fixated on the next glimpse. The systematic search routine also simulates the tendency of observers to eventually re-fixate objects that were previously fixated and found not to be targets. Fixation probabilities that were initially high and decreased tend to recover (increase again) after a number of glimpses. The recovery time depends on the number of blobs in the field of view. This is consistent with empirical studies of visual search.

3.3. Selective Attention and Perceptual Learning

There are at least two aspects of attention that are important to STA performance. One – the mechanism that determines eye fixations and preattentive shifts in visual attention – was discussed in the last section. A second concerns the nature of the visual features that contribute to preattentive "pop-out" of objects and whether those features are subject to modification through learning. In the 1970's, Ann Treisman and her colleagues argued that preattentive processing and selection occur only for objects that are uniquely distinguished by a single perceptual dimension, such as size, color, shape, and luminance. However, Jeremy Wolfe and his colleagues later showed that given sufficient practice, observers could preattentively identify

objects based on conjunctions of perceptual dimensions (e.g., find the red circle in a background of blue circles, blue squares, and red squares). Neisser and others have shown that, given enough experience with the stimulus, observers can reach a point where complex combinations of features support pop-out. For example, Neisser found that after extensive training, observers can learn to rapidly pick out a target letter that is very similar to the background clutter, e.g., a "K" in a field of Es, Hs, Ts, Ls, and Fs. Schneider and his colleagues have studied the development of "automaticity" or preattentive processing in letter search tasks. They showed that letters that are consistently "mapped" as one of a set of targets (as opposed to sometimes being targets and sometimes distractors) eventually become automatically processed after extensive practice.

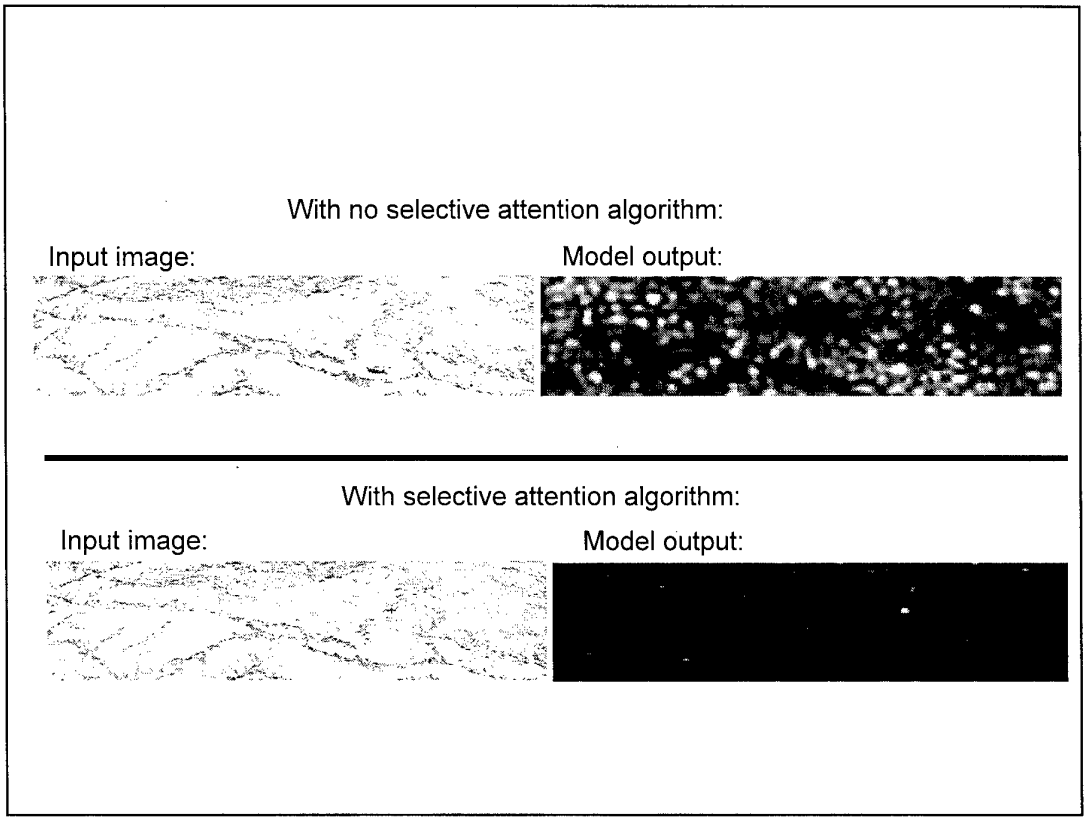


Figure 5. Clutter rejection performance of the GTV model.

After extensive practice, military observers are often able to immediately pick out targets in cluttered scenes that novice observers must search for painstakingly. They have evidently learned to preattentively process the target. It is therefore important to model the effect of learning on pop-out and visual search performance. One way of doing this is to differentially weight the filter-channel outputs before pooling them into a single salience map. The weights would be designed to amplify channel outputs typical of the target, and attenuate channel outputs typical of background clutter. The GTV model uses this method, employing a discriminant analysis routine to compute the weights. The weighting routine is highly effective in rejecting clutter, as shown in Figure 5, and it allows the model to simulate the performance of experienced human observers.

3.4. Other Cognitive Processes that Affect STA

Another aspect of cognition that affects search and target acquisition is perceptual decision making. Target acquisition is not simple signal-to-noise ratio threshold process, but involves decision-making. Signal detection theory describes observers' ability to trade-off detections versus false alarms. These trade-offs can distort the relative probabilities of detection in task of differing difficulty[13]. For example, we have previously reported that human observers tend to shift their decision criterion as the difficulty of the detection task changes. For example, in low clutter conditions the observer may adopt a relatively high decision likelihood ratio criterion, β. But when faced with high clutter, the same observers tend to relax β. This has the effect of allowing them to increase their probability of detection at the cost of a higher false alarm probability, as illustrated in Figure 6. This perceptual decision tradeoff process can have considerable impact on measured probabilities of detection.

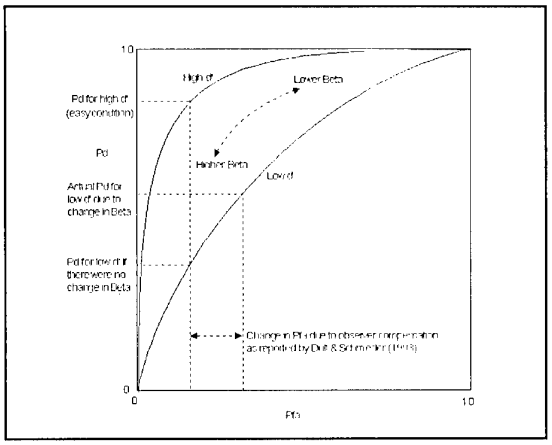


Figure 6. ROC curve showing shift in observer criterion with task difficulty.

The GTV model uses signal detection theory in two ways. When there are multiple "blobs" or areas of interest in the field of view, a decision must be made as to which blob the eyes will saccade to next. The extreme detector model is used to make this decision. The choice of blobs for the next saccade is highly non-linear -- even though one blob may have just slightly greater spatio-temporal contrast energy than the others, it's probability of fixation will be much larger. The metric used to describe each blob is actually a power function of its spatio-temporal contrast energy.

The GTV model also uses signal detection theory to decide whether or not the blob currently being fixated is a target. The spatio-temporal contrast metric for the current blob is compared to the distributions of the same metric for targets and clutter objects encountered during training. It is predicted that the observer says "yes, the blob is a target" when the ratio of the probability densities of the target to clutter distributions exceeds the criterion value of the decision likelihood ratio, β.

4. REQUIREMENTS FOR TESTING AND VALIDATING STA MODELS

There are a number of requirements for the design and conduct of successful validation tests that derive from an understanding of human vision and visual cognition. Although many investigators will be familiar with these requirements, one or more of the requirements have not been met in almost all STA model validation efforts. Exposition and discussion of requirements can therefore benefit the STA community.

- Since the sensitivity of the human visual system depends on the luminance level and chromaticity of the input scene, input images must be photometrically and colorimetrically calibrated. Some issues of color calibration are discussed by Rogers and Thomas[14].
- Since the human visual system has high acuity only in small portion of the visual field (i.e., the fovea), the likelihood that a target is foveated is an important determinant of overall detection probability. The larger the observer's field of view, the less likely it is that any given target will be foveated (assuming constant magnification). It is therefore important that the apparent field of view (AFOV) of the imagery used to test models be the same as the AFOV that observers used in the experiment whose results are to be matched.
- Simply instructing observers to make their responses indicative of a given level of processing (e.g., locate "areas of interest" without full detection or recognition) does not guarantee that they limit their processing to that level. If the observers are given enough time, they generally perform higher levels of processing (e.g., recognition or identification) before reporting the location of an area of interest. Even if exposure time is limited, observers may perform additional processing on the persisting iconic memory of the target. Observer validation experiments should therefore use brief image exposures followed by a noise mask pattern in order to limit processing.
- If the model under test requires training, the target and background images given the model during training must adequately sample the same target and background features that will be present in the test imagery.
- Two possible scenarios must be considered in determining the spatial resolution of imagery used to test a model: (a) the resolution in the observer test was limited by a display and/or sensor, or (b) the resolution in the observer test was limited only by the human eye, e.g., observers viewing targets with the naked eye or DVO in clear conditions. In the first case, the images submitted to the model must be filtered to simulate the MTF of sensor/display system. In the second case, the images provided as inputs to the model must have resolution at least as great as that of the human visual system. They must therefore be captured by a sensor whose resolution exceeds that of the human visual system.
- The temporal up-date rate of the imagery should be at least the Nyquist frequency of the highest rate of temporal modulation in the scene. Alternatively, frame rate can be set to the highest temporal cut-off frequency of the human eye. This last quantity will depend on the intensities, spatial frequencies, and chromaticities in the scene and the viewing conditions.

It should be noted that these requirements are a product of the complexity of human observers' visual performance -- not a consequence of the complexity of any model. They therefore apply regardless of whether one is testing a simple or a complex model.

5. ALTERNATIVE APPROACHES FOR FIELD TESTING AND MODEL VALIDATION

The process of validating search and detection models or metrics is expensive and time-consuming. It is therefore worth considering some of the alternative approaches available and the advantages and potential pitfalls of each. We contrast three different approaches here, all of which involve collection of imagery from the field and psychophysical tests with human observers in either the field or a laboratory. The three approaches are summarized in Table 1.

Table 1. Alternative approaches for field testing and STA model validation.

Approach A Observer test in the field	• Collection of imagery of targets in backgrounds, ground truth, ambient illumination, meteorological data, and calibration data in field with high resolution camera • Observer test in field viewing targets through DVO device • Field imagery and calibration data submitted to STA model to generate predictions • Model predictions compared to observer performance in field
Approach B Observer test in the laboratory with field imagery	• Collection of imagery of targets in backgrounds, ground truth, ambient illumination, meteorological data, and calibration data in field with high resolution camera • Observer test in the laboratory by displaying imagery from field test • Field imagery and calibration data submitted to STA model to generate predictions • Model predictions compared to observer performance in laboratory
Approach C Use and validation of synthetic imagery Observer test in the laboratory with synthetic imagery	• Collection of imagery of background only, ground truth, ambient illumination, meteorological data, and calibration data in field with high resolution camera • Measurement of Bi-directional Reflectivity Distribution Function (BRDF) of target paints • Synthetic target generated and inserted in calibrated background imagery from field • Synthetic imagery validated by comparing it to field imagery • Observer test in the laboratory with validated synthetic imagery • Synthetic imagery submitted to STA model to generate predictions • Model predictions compared to observer performance in laboratory

The conventional and most obvious approach is to collect both observer data and imagery to submit to the STA model in the field. This is Approach A in Table 1. One of the major disadvantages of Approach A is that it is difficult to control observer tests in the field. The field of view, exposure time, time of day, and cloud shadows experienced by observers all must be the same as those in the imagery collected for submission to the STA model. Moreover, the observers must be shielded from acoustic and social cues that would affect their STA performance. Another serious problem with Approach A is that no camera can reproduce the full range of colors and intensities that the observers experience in the field. Very high signals (e.g., from specular reflections) will exceed the dynamic range of the camera (i.e., saturate). If the camera gain is set lower, then low signals (e.g., in shadowed areas of the scene) will fall below the sensitivity threshold of the camera and these areas will appear black in the image.

One possible solution to these problems is to do the observer testing in the laboratory using imagery of targets in backgrounds collected from the field. This is Approach B in Table 1, and the approach used by TNO for the DISTAFF data set. This approach does not eliminate the camera dynamic range problem, but ensures that both the observers and the STA model are subject to the same effects in this regard. However, this approach still suffers from other disadvantages (which are also present in Approach A). For one, it is expensive and time-consuming to deploy real targets in the field in a controlled manner. The very act of deploying them also produces extraneous detection cues, such as vehicle tracks.

Capturing temporal effects is also a problem in both approaches A and B. If one wants to capture important effects of target motion (relative to the background, or motion of parts of the target relative to the whole), then the problems of field deployment and control are compounded. For example, the rate and pattern of motion of a vehicle over rough terrain may be an important detection/recognition cue. Shadows produced by clouds and the motion of helicopter rotor blades are other temporal effects that can greatly influence detection. Capturing these motion effects in imagery requires a very high frame rate, and results in a huge amount of imagery that must be stored and calibrated.

Extraneous cues from target deployment can be eliminated and temporal effects controlled by using synthetic imagery for both the observer test and as input to the STA model. This approach is used in the VISEO system[15], and is shown as Approach C in the above table. This is a two-step approach – first synthetic imagery is generated and validated, and then the STA model is validated using the

synthetic imagery. The VISEO system generates backgrounds using one or more spectral bands of measured background imagery, depending on the type of sensor being simulated. The spectral bands range from the visible to LWIR. The database is calibrated, and the algorithm for combining bands has been validated.[16] The VISEO system also has a library of approximately 75 high-fidelity ground and air targets, most of which have been validated in the visible and/or IR bands. With the VISEO system, one can generate imagery at any desired frame rate in order to capture high temporal-frequency effects. With VISEO, one need not generate the imagery for the whole set of test conditions at one time. Imagery can be generated for selected conditions, submitted to the STA model to generate predictions, and then archived. Another advantage of the VISEO system is that radiation from the target model is not limited by any camera or sensor system. One can therefore model specular reflections from the target, for example, and evaluate their effect on detectability by submitting the resulting scene data directly to the STA model.

It is clear that Approach A has serious shortcomings – due both the difficulty of controlling observer test in the field and sensor dynamic range limitations. However, approaches B and C both have advantages for certain types of applications. With VISEO, there is no need to deploy and control targets during field imagery collection, and one can more easily evaluate temporal effects and specular reflections. However, one must build high-fidelity models of the targets, if they are not already in the VISEO database.

6. REFERENCES

1. Doll, T. J., McWhorter, S. W., Wasilewski, A. A. and Schmieder, D. E., "Robust, sensor-Independent target detection and recognition based on computational models of human vision", *Optical Engineering*, 37, (7), pp. 2006-2021, 1998.
2. Walker, G.W., and McManamey, J.R., "The importance of second-order statistics for predicting target detectability", In: *Proceedings SPIE*, vol. 1967, pp. 308-319, 1993.
3. Rotman, S. R., Hsu, D. Cohen, A., Shamay, D., and Kowalczyk, M. L., "Textural metrics for clutter affecting human target acquisition", *Infrared Physics & Technology*, 37, pp. 667-674, 1995.
4. Northdurft, H. C., "Texture segmentation and pop-out from orientation contrast", *Vision Research*, 31, pp. 1073-1078, 1991.
5. Gould, J. D., "Looking at pictures", In: R. A. Monty and J. W. Senders (Eds.), *Eye movements and psychological processes*, Hillsdale, N. J.: Lawrence Erlbaum Associates, 1976.
6. Rayner, K., "Foveal and parafoveal cues in reading", In: J. Requin (Ed.), *Attention and performance* (Vol. 7), Hillsdate, NJ: Lawrence Erlbaum and Associates, 1978.
7. Bellamy, L. J. and Courtney, A. J., "Development of a search task for the measurement of peripheral visual acuity", *Ergonomics*, 24, pp. 497-509, 1981.
8. Erikson, R. A., "Relation between visual search time and visual acuity", *Human Factors*, 6, pp. 165-178, 1964.
9. Johnston, D. M., "Search performance as a function of peripheral acuity", *Human Factors*, 7, pp. 528-535, 1965.
10. Arani, T., Karwan, M. H. and Drury, C. G., "A variable-memory model of visual search", *Human Factors*, 26, pp. 631-639, 1984.
11. Humphereys, G.W. and Muller, H.J., "Search via Recursive Rejection (SERR): A Connectionist Model of Visual Search", *Cognitive Psychology*, 25, pp. 43-110, 1993.
12. Nicoll, J.F., *A Search for Understanding: Analysis of Human Performance on Target Acquisition and Search Tasks Using Eyetracker Data*, IDA Paper P-3036, Institute for Defense Analyses, 1995.
13. Doll, T. J. and Schmieder, D. E., "Observer false alarm effects on detection in clutter", *Optical Engineering*, 32, pp. 1675-1684, 1993.
14. Rogers, G.A. and Thomas, D.J., *Improved color image calibration*, Report ???, U.S. Army Combat Systems Test Activity, Aberdeen Proving Ground, MD. U.S. Army Tank-Automotive Research Development and Engineering Center, 19??
15. Doll, T. J., McWhorter, S. W., Schmieder, D. E., Hetzler, M. C., Stewart, J. S., Wasilewski, A. A., Owens, W. R., Sheffer, A. D. Galloway, G. L., and Harbert, S. D., *"VISEO" System for Camouflage /LO Design based on Computational Vision Research*, In: Proceedings of the HAVE FORUM Low Observables Symposium, Eglin AFB, FL, Report No. WL-TR-97-6003, 1997.
16. Doll, T. J., McWhorter, S. W., Hetzler, M. C., Stewart, J. M., Wasilewski, A. A., Schmieder, D. E., Owens, W. R., Sheffer, A. D., Galloway, G. L., and Harbert, S. L., *Visual / Electro-Optical (VISEO) Detection Analysis System: Final Report*, Report No. USAAMCOM TR 98-D-19, Atlanta, GA: Georgia Tech Research Institute, 1997.

VISUAL DISTINCTNESS DETERMINED BY PARTIALLY INVARIANT FEATURES.

J.A. Garcia, J. Fdez-Valdivia
Departamento de Ciencias de la Computacion e I.A.
Univ. de Granada. E.T.S. de Ingenieria Informatica.
18071 Granada. Spain
E-mail: jags@decsai.ugr.es, J.Fdez-Valdivia@decsai.ugr.es

Xose R. Fdez-Vidal
Departamento de Fisica Aplicada.
Univ. de Santiago de Compostela. Facultad de Fisica.
15706 Santiago de Compostela. Spain
E-mail: faxose@usc.es

Rosa Rodriguez-Sanchez
Departamento de Informatica. Universidad de Jaen
Escuela Politecnica Superior. 23071 Jaen. Spain
E-mail: rosa@ujaen.es

1. SUMMARY

This paper describes a system for the automatically learned partitioning of "visual patterns" in digital images, based on a sophisticated, band-pass, filtering operation, with fixed scale and orientation sensitivity. The "visual patterns" are defined as the features which have the highest degree of alignment in the statistical structure across different frequency bands. Here we show a computational visual distinctness measure computed from the image representational model based on visual patterns. It is applied to quantify the visual distinctness of targets in complex natural scenes. We also investigate the relation between the computational distinctness measure and the visual target distinctness measured by human observers.

2. INTRODUCTION

Images issued from the environment should not be presumed to be random patterns. Instead, real-world images contain characteristic statistical regularities that set them apart from purely random images. There are a number of statistical properties that we might consider when looking at real-world images, and many of the important forms of structure that are contained in 2D images require higher-order statistics characterization. Moreover, Field [1] noted that there is likely to be a variety of features which extend across different frequency bands. For instance, the presence of edges and lines in an image corresponds to a type of congruence between the different scales of the image which is destroyed when the phases are randomized [2]. These features exist because some degree of alignment exists between the phases at different frequencies. There are also other forms of congruence across scales in 2D digital images. Field [3] suggested that the power spectra of natural images falls off as a function of frequency by a factor of approximately $1/k^2$. This implies that the image will have constant variance across scales: the contrast as measured by the variance in pixel intensities should remain roughly constant, independently of the viewing distance.The perceptual organization capabilities of human vision seem to exhibit the properties of detecting viewpoint-invariant structures and calculating varying degrees of significance for individual instances [4]. Lowe [5] proposed that the structures to be detected in the image should be formed bottom-up using perceptual grouping operations that exhibit exactly these properties in the absence of domain knowledge, yet must be of sufficient specificity to serve as indexing terms into a database of objects. Given that we often have no priori knowledge of viewpoint for the objects in a database, these indexing features that are detected in the image must reflect properties of the objects that are at least partially invariant over a wide range of viewpoints of some corresponding three-dimensional structure. This means that it is useless to look for features with particular sizes or orientations or other properties that are highly dependent upon viewpoint. The second constraint on these indexing features is that there must be some way to distinguish the relevant features from the dense background of other image features which could potentially give rise to false instances of the structures.

Often implicit in the interpretation of visual search tasks is the assumption that the detection of targets is determined by the feature-coding properties of low-level visual processing [6]. Instead of assuming that perceived shapes are simple or statistical structure at a particular scale, we think it more appropriate to regard them as "visual patterns", distinguished at an object level.

Here we show a particular scheme for filtering observed images, designed to the automatically learned partitioning of features (visual patterns) which have the highest degree of alignment in statistical structure across different frequency bands. These features are likely to be invariant over a range of scales and orientations and can be judged unlikely to be accidental in origin even in the absence of specific information regarding which objects may be present. Then, we present a computational visual distinctness measure computed from the image representational model based on visual patterns. This measure applies a simple decision rule to the distances between segregated visual patterns, and it will be used to quantify the visual distinctness of targets in complex natural scenes. The analysis to the automatically learned partitioning of "visual patterns" (it has been termed RGFF model) follows three stages: Preattentive stage, Integration stage, and Learning stage. Fig. 1 shows a general diagram describing how the data flows through the RGFF model. This diagram illustrates the analysis on a given image of a military vehicle in a complex rural background.

In the preattentive stage of the RGFF system (Section 3), the clumps of energy in the Fourier spectrum of the image are captured into a collection of oriented spatial-frequency channels, as illustrated in Fig. 1. The segregation of these clumps of energy induces the selection of a subset of activated filters (which are selectively sensitive to them) from a filter bank of log-Gabor functions centered at 12 orientations and 5 ranges. Due to conjugate symmetry, the filter design is only carried out on half the 2D frequency plane. The activated log-Gabor filters produced by the preattentive stage are illustrated in the diagram by ellipses drawn, in the 2D spatial-

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

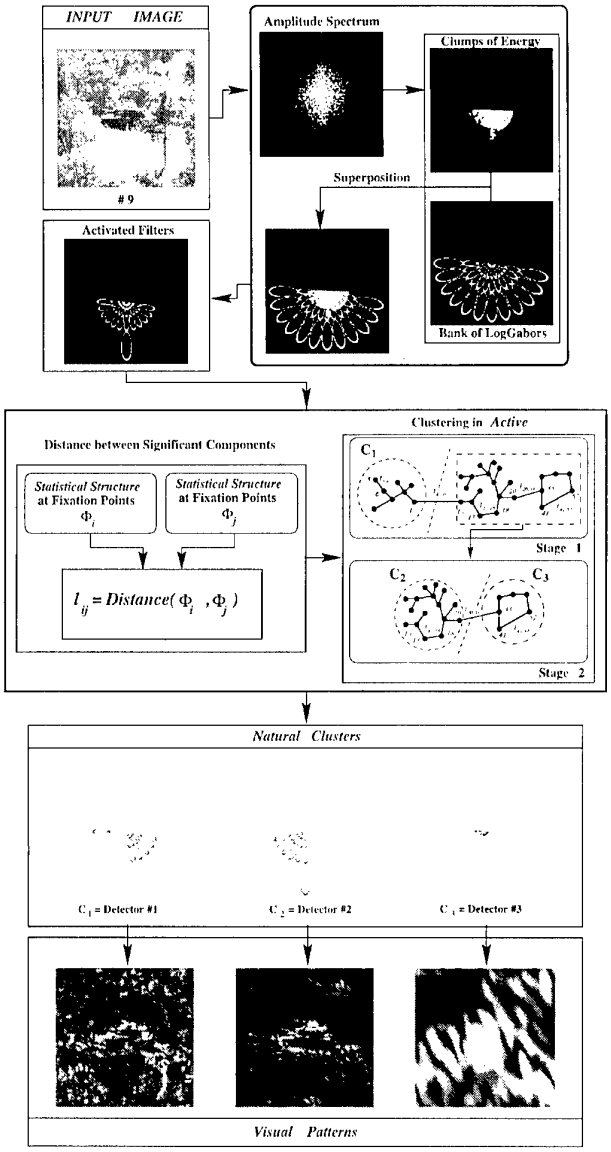


Figure 1: A general diagram describing how the data flows through the representational model..

frequency plane, at the point where their amplitude has decreased to the ($e^{-1/2}$) half width its maximum.
In the integration stage (Section 4), for any two activated filters, their responses are compared based on the distance (a β-norm between their statistical structure, computed over those pixels which form "fixation points" of the filters (local energy peaks on the filtered response).
In the learning stage (Section 5), clustering on the basis of the distance between the activated filters is performed to highlight scale and orientation invariance of responses.
As shown in Fig. 1, three collections of filters were obtained in the Learning stage for the input image in accordance with a constraint of invariance in statistical structure across frequency bands. The filtered responses of activated log-Gabors in each one of the three groupings were summed for the automatic learned partitioning of the visual patterns.
The performance of this notion of visual pattern to segregate potential targets can be visually evaluated in Fig. 1, at the bottom. The dominant signal in the output from detector #2 is the military vehicle (target) which is well preserved. On the contrary, both large structures and fine detail of the natural background were removed, even though significant background clutter that can affect the target distinctness is still present. In fact, the fine details of the natural background, which are not significant for quantifying the target distinctness, are isolated in the output from detector #1. And the lower frequency texture of the background is segregated into the output from detector #3. Fig. 2 demonstrates the ability of the same model to achieve signal separation from superposition of objects on three synthetic images. The image in Fig. 2.A1 was partitioned into two "visual patterns", as shown in Figs. 2.C1 and 2.E1. In the learning stage, the set of activated filters was partitioned into two groupings of filters, as shown in Figs. 2.B1 and 2.D1. The "visual pattern" shown in Fig. 2.C1 (respectively, Fig. 2.E1) was obtained by the sum of the responses over filters in Fig. 2.B1 (resp., Fig. 2.D1). The "visual patterns" obtained by the model on Fig. 2.A2, are illustrated in Figs. 2.C2 and 2.E2. The learning stage produced two collections of filters as shown in Figs. 2.B2 and 2.D2. The right column in Fig. 2 shows the signal separation achieved by the analysis on the input image given in Fig. 2.A3.
Finally, Section 6 presents the computational visual distinctness measure computed from the image representational model based on visual patterns. As illustrated in Fig. 10, this measure applies a simple decision rule to the distances between segregated visual patterns.

3. PREATTENTIVE STAGE

In the RGFF model, the encoding strategy will rely on the combined activity of subsets of filters. Only a small number of units will contribute to the detection of each visual pattern. These collections of filters will be derived from a learning stage, based on the degree of congruence between the responses of strongly responding filters that a preattentive stage produces. There are two basic assumptions for this first stage:

1. Spatial information on the image is analyzed by multiple filters, each of which is sensitive to patterns whose spatial frequencies are in a particular range.
2. The RGFF model bases its responses only on those filters sensitive to relevant forms in the complex scene.

These assumptions are in agreement with models of spatial-frequency channels which are quite successful for the detection of visual patterns [7]. The output of the preattentive stage will be the units from a fixed filter bank of log-Gabors which are tuned to the clumps of energy in the Fourier spectrum of the given image. The selected units are the filters in the bank which strongly respond to some pattern that the image contains. These filters are referred as the "activated" filters of the bank. Also for each activated filter, pixels whereupon the focus of attention should be shifted to measure congruence and which form "fixation points", are computed as local energy peaks on the filtered response. This processing is based on current models of human visual search and detection which assume that a preattentive stage indicates potentially interesting image regions, and where a serial stage is deployed to analyze them in detail [6,7].

3.1. Bank of filters

The set of filters used in the decomposition of the picture consists of log-Gabor filters of different spatial frequencies and orientations [3]. Log-Gabor functions, by definition, have no DC component. The transfer function of the log-Gabor has extended tails at the high frequency end. Thus it should be able to encode natural images more efficiently than ordinary Gabor functions, which would over-represent the low-

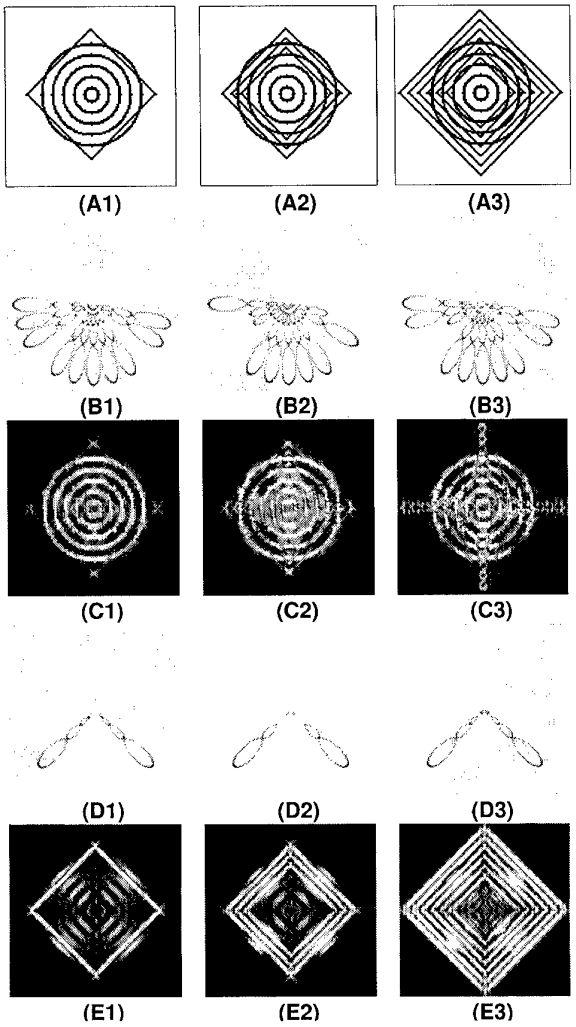


Figure 2: Automatically learned partitioning of "visual patterns" in synthetic image data.

frequency components and under-represent the high frequency components in any encoding process. Another argument in support of the log-Gabor functions is the consistency with measurements on the mammalian visual system [8].

A Log-Gabor filter determines a Gaussian in the spatial frequency domain around some central frequency (r_o, θ_o). It can be represented in the frequency domain as the sum of the even-symmetric log-Gabor filter and i times the odd-symmetric log-Gabor filter as follows:

$$\phi(r_o,\theta_o) = \exp\left\{-\frac{(\log(r/r_o))^2}{2(\log(\sigma_r/r_o))^2}\right\}\exp\left\{-\frac{(\theta-\theta_o)^2}{2\sigma_\theta^2}\right\} \qquad (1)$$

where θ_o is the orientation angle of the filter, r_o is the central radial frequency, σ_θ and σ_r are the angular and radial sigma of the Gaussian, respectively.

The convolution of a log-Gabor function (whose real and imaginary parts are in quadrature) with a real image results in a complex image. Its norm is called energy and its argument is called phase. The local energy of the image analyzed by a log-Gabor filter (hereafter, filtered response) can be expressed as [3]:

$$E(x,y) = \sqrt{O_{even}^2(x,y) + O_{odd}^2(x,y)} \qquad (2)$$

where $O_{even}(x,y)$ is the image convolved with the even-symmetric log-Gabor filter and $O_{odd}(x,y)$ is the image convolved with the odd-symmetric log-Gabor filter.

The real-valued function given in equation (1) can be multiplied by the frequency representation of the image and after transforming the result back to the spatial domain, the results of applying the oriented energy filter pair are extracted as simply the real component for the even-symmetric filter and the imaginary component for the odd-symmetric filter [9]. The bank of the filters should be designed so that it tiles the frequency plane uniformly (the transfer function should be a perfect bandpass function). The length to width ratio of the filters controls their directional selectivity. The ratio can be varied in conjunction with the number of orientations used in order to achieve a coverage of a 2D spectrum. Furthermore, as the degree of blurring introduced by the filters increases with their orientational selectivity, they must be carefully chosen to minimize the blurring. Hence we consider a filter bank with the following features:

1. The spatial frequency plane is divided into 12 different orientations.
2. The radial axis is divided into 5 equal octave bands. In a band of width 1 octave, spatial frequency increases with a factor 2. The highest filter (for each direction) is positioned near the Nyquist frequency to avoid ringing and noise. The wavelength of the five filters in each direction is set at 3, 6, 12, 24, and 48 pixels, respectively.
3. The radial bandwidth is chosen as 1.2 octaves.
4. The angular bandwidth is chosen as 15 degrees.

Twelve different angles for each resolution are chosen and five different resolutions are used. The resultant filter bank is illustrated in Fig.1. Due to conjugate symmetry, the filter design is only carried out on half the 2D frequency plane. The log-Gabor filters are illustrated in the diagram by ellipses drawn, in the 2D spatial-frequency plane, at the point where their amplitude has decreased to the $(e^{-1/2})$ half width its maximum.

3.2. Activated filters in the bank

In order to decompose the image into its most significant components, strongly responding filters should be selected for the input image.

Let *Active* be the set of filters in the bank that strongly respond to the spatial information content. They will be selectively sensitive to patterns in the scene. These patterns produce clumps of energy upon the Fourier spectrum of the image. The activated units from the bank are then simply those filters whose amplitude spectrum and some clump of energy in the image amplitude spectrum overlap to some extent, as illustrated in Fig.1.

3.3. Selection of fixation points

In the integration stage, for any given two activated filters, a distance between them is derived via distances between their statistics. The distance chosen is the β-norm, computed over those pixels which form "fixation points" of the filters. The "fixation points" are simply local energy peaks on the filtered response. The standard argument for selecting regions of high Gabor energy is that they would provide a good starting point for exploring common grounds between several activated filters in the Gabor space. The implementation of the local-energy model used here is the one presented in [10]. Given the original image, the local energy map E_i for the activated filter ϕ_i, given in equation (2), yields a representation in the

space spanned by two functions, $O_{even}(x,y)$ and $O_{odd}(x,y)$, where $O_{even}(x,y)$ is the image convolved with the even-symmetric log-Gabor filter and $O_{odd}(x,y)$ is the image convolved with the odd-symmetric log-Gabor filter at (x,y). Hence, the detection of peaks on the E_i map acts as a detector of significant features on the filtered response.

4. INTEGRATION STAGE

Given a decomposition of the original image into its most significant components, only a further element is needed to define the concept of visual pattern: a distance measure, denoted as *Distance*(ϕ_i, ϕ_j) , between the statistical structures of the filtered responses for each pair of filters ϕ_i and ϕ_j (Section 4.2.). Then, *Distance*(ϕ_i, ϕ_j) returns a value of the degree of congruence between statistical structure at different scales and orientations.

There are two basic assumptions for measuring congruence between two filtered responses in this second stage:

1. The similarity between two filtered responses can be measured by the Quick pooling of the differences between their statistical structure.
2. The measure of similarity is not simply computed globally over the entire filtered response, but semi-locally at locations that are local energy peaks (fixation points).

Previously, it was demonstrated [11] that a measure based on these two assumptions produces a good predictor of target saliency for humans performing visual search and detection tasks.

4.1. Definition of integral feature for the partitioning of visual patterns

For each activated filter ϕ_i, the respective filtered response may be represented by any subset of the following separable features:

1. The phase value defined as:

$$T_1^i(x,y) = arctan\frac{O_{even}(x,y)}{O_{odd}(x,y)} \tag{3}$$

where $O_{even}(x,y)$ is the image convolved with the even-symmetric log-Gabor filter of ϕ_i , and $O_{odd}(x,y)$ is the image convolved with the odd-symmetric log-Gabor filter of ϕ_i at (x,y) (Section 3.1.).

2. A normalized measure of local energy as given by:

$$T_2^i(x,y) = \frac{E_i(x,y)}{\sum_{\{j/\phi_j \in Active\}} E_j(x,y)} \tag{4}$$

where $E_i(x,y)$ denotes the local energy at (x,y) for filter ϕ_i (see equation 2 for further details), and *Active* is the set of activated filters for the image. This definition of a normalized local energy incorporates lateral interactions among activated filters to account for between-filter masking.

3. The local standard deviation of the normalized local energy defined as:

$$T_3^i(x,y) = \left(\frac{1}{Card\ [W(x,y)]} \sum_{(p,q)\in W(x,y)} (T_2^i(p,q) - \mu)^2 \right)^{1/2} \tag{5}$$

where

$$\mu = \frac{1}{Card\ [W(x,y)]} \sum_{(p,q)\in W(x,y)} T_2^i(p,q)$$

and $T_2^i(p,q)$ as given in equation (4). The neighborhood $W(x,y)$ is defined as the set of pixels contained in a disk of radius r centered at (x,y). Let r be defined as the Euclidean distance between (x,y) and the nearest local minimum to (x,y) on the energy map E_i. Since the nearest local minimum to (x, y) on the local energy map marks the beginning of another potential structure, our selection for the neighborhood $W(x, y)$ avoids interference with such a structure while the local variation is computed [10].

4. The local contrast of the normalized local energy defined as:

$$T_4^i(x,y) = \frac{T_3^i(x,y)^2}{\mu} \tag{6}$$

where

$$\mu = \frac{1}{Card\ [W(x,y)]} \sum_{(p,q)\in W(x,y)} T_2^i(p,q)$$

5. The local entropy of the normalized local energy within $W(x,y)$, noted as $T_5^i(p,q)$.

Although we propose these five features, any other intent to capture relevant characteristics of the scene, while stable for the representation of the image is also conceivable. Hereafter, *an "integral feature" is defined as a particular subset of separable features at a fixation point* [12].

For representing the filtered responses of the input image, different definitions of integral feature can be given based on different subsets of separable features. Consequently, the system should learn the best integral feature definition for the input image in which to look for invariance across orientations and scales. This point is analyzed in Section 6.4.

4.2. Congruence in integral features between two filtered responses

In order to define a distance between the integral features of two filtered responses, we need to specify how the differences in each separable feature are to be pooled into an overall difference at fixation points.

Let ϕ_i and ϕ_j be a pair of activated filters in *Active*. Let $T^i(x,y)=(T_{l_k}^i(p,q))_{1\le k\le L}$, with $l_k \in \{1,2,...,5\}$, be the integral feature at (x,y) computed on the filtered response of ϕ_i, based on a number of L separable features (Section 4.1.). In a similar way, let $T^j(x,y) = (T_{l_k}^j(p,q))_{1\le k\le L}$, be the integral feature at (x,y) on the filtered response of ϕ_j.

We take $D[T^i(x,y), T^j(x,y)]$ defining a distance measure between integral features $T^i(x,y)$ and $T^j(x,y)$ as given by the equation:

$$D[T^i(x,y),T^j(x,y)] = \sum_{k=1}^{L} \frac{1}{Max_{l_k}} d(T_{l_k}^i(x,y),T_{l_k}^j(x,y)) \tag{7}$$

where normalization Max_{l_k} is defined as:

$$Max_{l_k} = \max_{m,n,\phi_m,\phi_n \in Active} \left\{ d(T_{l_k}^m(p,q), T_{l_k}^n(p,q)) \,/\, (p,q) \in FP(n) \right\}$$

with $FP(n)$ being the fixation points for the activated filter ϕ_n and *Active* being the set of activated filters; and where for $l_k=1$, we have:

$$d(T_1^i(x,y),T_1^j(x,y)) = \left| arctan\frac{sin(T_1^i(x,y)-T_1^j(x,y))}{\cos(T_1^i(x,y)-T_1^j(x,y))} \right| \tag{8}$$

and for $l_k = 2,3,4,5$:

$$d(T_{l_k}^i(x,y),T_{l_k}^j(x,y)) = \left| T_{l_k}^i(x,y) - T_{l_k}^j(x,y) \right| \tag{9}$$

The congruence in integral features between two filtered responses is computed by using Quick pooling [13]. It is the most common model of integration over spatial extent, and is

essentially the square root of the squares sum except that the exponent is not restricted to the value of 2. The Quick pooling can be viewed as a metric in a multidimensional space, and it is sometimes known as Minkowski metric.

The distance between the filtered responses of ϕ_i and ϕ_j, which provides a measure of the extent to which features extend through frequency, is given by:

$$Distance(\phi_i,\phi_j)=Dist^2[i,j]+Dist^2[j,i] \tag{10}$$

where:

$$Dist[p,q]=\frac{1}{Card[FP(p)]}\left(\sum_{(x,y)\in FP(p)}\left|D[T^p(x,y),T^q(x,y)]\right|^\beta\right)^{\frac{1}{\beta}} \tag{11}$$

with $FP(p)$ being the set of fixation points for the activated filter ϕ_p; and where $D[T^p(x,y), T^q(x,y)]$ is defined as given in equation (7).

The default value of the exponent β in equation (11) is 3. Graham [7] discussed at some length several interpretations of the Quick pooling formula and the selection of the pooling exponent.

5. LEARNING STAGE

Based on a measure of the extent to which features extend through frequency, noted as Distance(ϕ_i,ϕ_j), a "visual pattern" is simply defined as congruence in statistical structure, as measured by Distance, across a range of 2D spatial frequency bands.

The individual filters spanning this particular range of bands will determine a natural cluster of units, noted as C_n, in the set of activated logGabors *Active*. By taking into account the statistical congruence across this range of frequency bands, a pair of filters ϕ_i and ϕ_j will belong to the same natural cluster C_n if there exists certainly continuity (i.e., there exists similarity in some statistics at the same spatial locations) across the filtered responses for an intermediate sequence of filters, between ϕ_i and ϕ_j, in C_n.

Therefore the definition of "visual pattern" induces a partition in *Active* into a number of natural clusters $C_1, C_2, ..., C_N$ such that:

$$Active=\bigcup_{n=1}^{N}C_n \text{ , and } C_p\cap C_q=\phi, \quad with \quad p\neq q, p,q=1,2,..,N \tag{12}$$

where, for each C_n, a pair of filters ϕ_i , $\phi_j \in C_n$ if there exists a sequence of filters $\phi_{n_1}, \phi_{n_2},..., \phi_{n_l}$ in C_n such that

$$\begin{aligned} &Distance\ (\phi_i,\phi_{n_1})\le\varepsilon_n \\ &Distance\ (\phi_{n_l},\phi_j)\le\varepsilon_n \\ &Distance\ (\phi_{n_k},\phi_{n_{k-1}})\le\varepsilon_n, k=1,2,...,l-1 \end{aligned} \tag{13}$$

where ε_n denotes the degree of statistical congruence between a pair of filters in C_n and verifies that:

$$\begin{aligned} &Distance\ (\phi_p,\phi_q)>\varepsilon_n, \\ &\forall\,\phi_p,\phi_q:\phi_p\in C_n,\phi_q\in Active-C_n \end{aligned} \tag{14}$$

The clustering of activated filters is performed as described in Section 5.1. Figs. 3 and 4 illustrate the performance of the clustering on several images of a target in a complex rural background.

The image in Fig. 3.A1 was partitioned into the two visual patterns shown in Figs. 3.C1 and 3.E1.

In the clustering process, the set of activated filters was partitioned into two collections of filters, as shown in Figs. 3.B1 and 3.D1. The "visual pattern" shown in Fig. 3.C1 (respectively, Fig. 3.E1) was obtained by the sum of the responses over filters in Fig. 3.B1 (resp., Fig. 3.D1).

The right column in Fig. 3 shows the separation achieved by the analysis on the image in Fig. 3.A2.

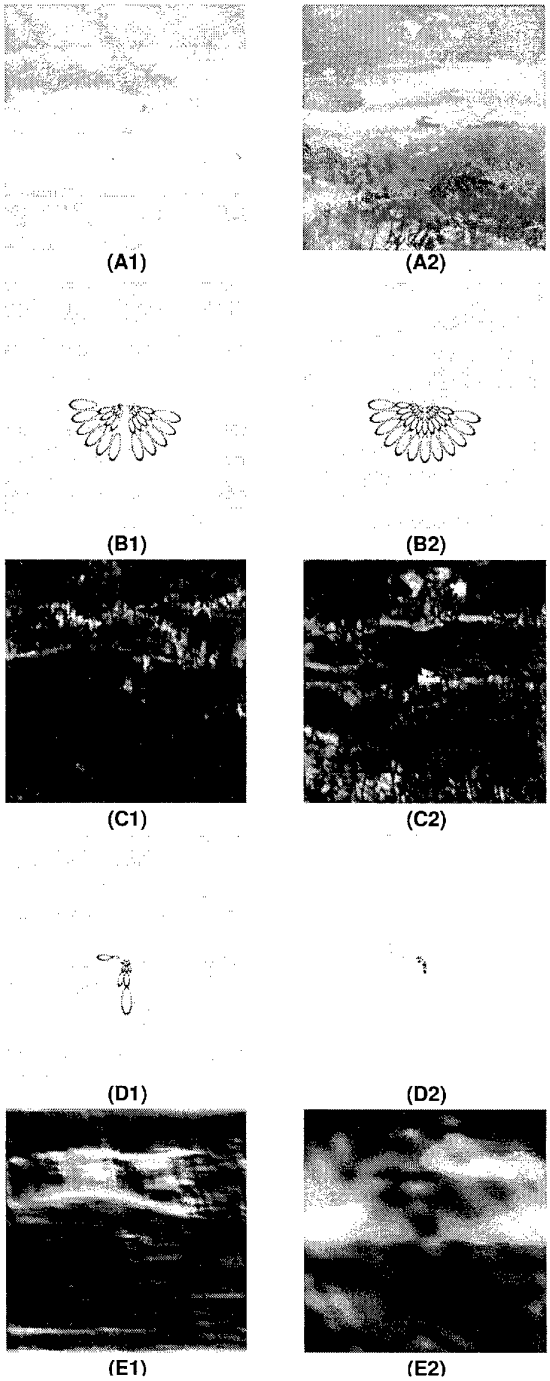


Figure 3: Natural clusters of activated filters and the respective visual patterns.

The visual patterns produced by the model on Fig. 4.A1 (respectively, 4.A2) are illustrated in Figs. 4.C1, 4.E1, and 4.G1 (resp., Figs. 4.C2, 4.E2, and 4.G2). In both cases, the clustering of activated filters produced three collections of filters as shown in Fig. 4.

5.1. Clustering of activated filters

We formulate the problem as the clustering of a dataset $X=\{i \mid \phi_i \in Active\}$ into a number N of natural clusters $\zeta_0, \zeta_1, ..., \zeta_{N-1}$. We call clusters natural if the membership is determined fairly well in a natural way by the data.

This clustering is reduced to a sequence of stages of simpler partitioning [14]. At each stage j, a subset X^j of X is divided into only two classes (for $j=0$, $X^0 = X$):
1. a natural cluster ζ_j which contains all the data points (filters) in X^j which are assigned the same class of a seed point (filter) $seed_j$, with $seed_j$ being picked randomly from X^j, and
2. the data, X^j-ζ_j, still not placed in any existing cluster, noted as $\zeta_0, \zeta_1, \ldots, \zeta_{j-1}$.

The clarity of separation between clusters, as measured by a dissimilarity function, will be the criterion by which we derive a natural cluster ζ_j at stage j. The dissimilarity function is defined in Section 5.1.1. The criterion by which we define a natural cluster at stage j is presented in Section 5.1.2. The dynamic process of clustering is stopped at stage j if the class X^j-ζ_j is the empty set. Otherwise, the process progresses, and the subset X^{j+1} to be partitioned at the stage $j+1$, it will be the one defined as $X^{j+1}=X^j$-ζ_j. Finally, the natural clusters in *Active* verifying equations (12)-(14) are induced as:

$$C_n = \{\phi_i \in Active \mid i \in \zeta_{n-1}\} \quad with \quad n = 1,2,\ldots, N \qquad (15)$$

and where N denotes the number of clusters into which $X=\{i \mid \phi_i \in Active\}$ was partitioned, that is $\zeta_0, \zeta_1, \ldots, \zeta_{N-1}$. See Fig. 1 for further illustration of this analysis.

5.1.1. Dissimilarity function

Let X^j be a subset of data not absorbed in any of the existing clusters $\zeta_0, \zeta_1, \ldots, \zeta_{j-1}$, at the stage j of the dynamic processing; with X^0 being the given data set. $X^0 = X = \{i \mid \phi_i \in Active\}$. Next we define a graph $GRAPH^j = (X^j, U^j)$ corresponding to the data subset X^j, and with U^j being the set of arcs $u=(k,l)$ between pairs of points in X^j. We associate with each arc $u \in U^j$ a real number $l(u) \geq 0$, and if $u=(k,l)$, we shall also use the notation l_{kl} for $l(u)$. Let l_{kl} be the distance from k to l defined as:

$$l_{kl} = Distance \quad (\phi_k, \phi_l) \qquad (16)$$

where $Distance(\phi_k, \phi_l)$ measures the distance between the filtered response of filters ϕ_k and ϕ_l as given in equation (10). The cost of a path is defined as the greatest distance between two successive vertices on the path. Let $\mu(seed_j, k)$ be a set of arcs constituting a path between two points $seed_j$ and k in X^j. And let $l(\mu)$ represent the cost of $\mu(seed_j, k)$ from $seed_j$ to k defined as follows:

$$l(\mu) = max\{l(u) \mid u \in \mu(seed_j, k)\} \qquad (17)$$

Taking into account that two filters belong to the same cluster if there exists continuity (i.e., there exists similarity in their statistics at the same spatial locations) across the responses of filters in a path between them, the dissimilarity function is next defined as the cost of the optimum path from a seed point to each other on the graph. The optimum path between two data points $seed_j$ and k is the path $\mu^*(seed_j, k)$ from $seed_j$ to k whose maximum *cost* $l(\mu^*)$ is minimum:

$$\mu^*(seed_j, k) = \underset{\mu}{Argmin} \left[max\{l(u) \mid u \in \mu(seed_j, k)\} \right] \qquad (18)$$

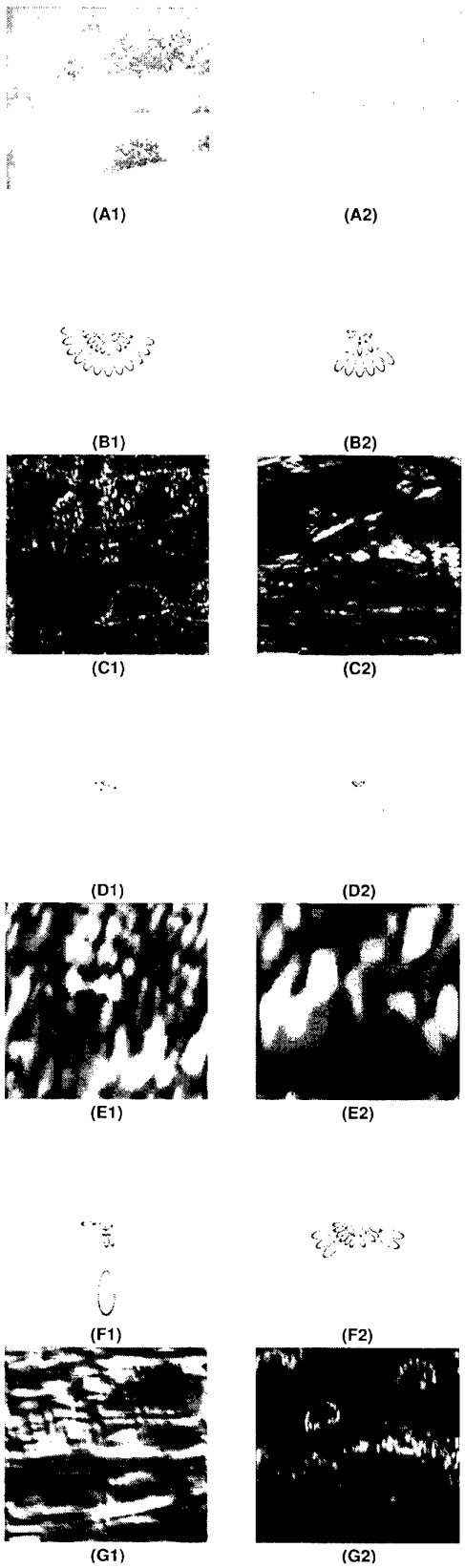


Figure 4: Natural clusters of activated filters and the respective visual patterns.

Visual target distinctness						
Image Pair	R_{P_d}		*RMSE*		VP_T	
	Value	Rank	Value	Rank	Value	Rank
# 16	96.39	1	2.55	4	4.799	1
# 9	96.27	2	16.90	1	4.531	2
# 37	96.19	3	16.12	2	3.815	3
# 6	90.95	4	1.94	7	3.637	4
# 30	90.66	5	2.37	5	2.713	6
# 26	90.06	6	1.71	8	2.969	5
# 29	89.47	7	1.60	9	2.234	9
# 3	80.02	8	2.83	3	2.254	8
# 21	73.84	9	1.35	10	2.046	10
# 11	73.40	10	1.96	6	2.690	7
P_{cc}	-		0.5		0.8	

Table 1: Comparative results of the RMSE metric and the computational visual distinctness measure.

Hence, the dissimilarity from the viewpoint of $seed_j$ to each k, is defined as the cost of the optimum path $\mu^*(seed_j,k)$ from $seed_j$ to each k:

$$d^{GRAPH^j}(seed_j,k)=l(\mu^*)=max\{l(u)|u\in\mu^*(seed_j,k)\} \quad (19)$$

with μ^* being the optimum path between $seed_j$ and k. The optimal path algorithm is given in [14].

5.1.2. Clarity of separation at stage j

Here we introduce the criterion by which we define the natural cluster ζ_j at stage j.

The set $\{d^{GRAPH^j}(seed_j, k), \text{with } k\in X^j\}$ is firstly ordered to obtain a new function:

$$d_j(i)=d^{GRAPH^j}(seed_j,k_i),\ such\ that\ d_j(i)\le d_j(i+1) \quad (20)$$

where $d_j(i)$ denotes the cost of the optimum path from $seed_j$ to k_i.

Let ε_j represent the degree of closeness that is required between a pair of points that belong to the natural cluster of $seed_j$, noted as ζ_j. Taking into account that $d_j(i)$ measures the closeness between $seed_j$ and k_i with $k_i \in X^j$, we have that ε_j can be defined as:

$$\varepsilon_j = d_j(i^*) \quad (21)$$

with i^* being the location of the first significant rise in the value of $d_j(i)$ when i increases. The value of i^* is computed as the first zero crossing of the second derivative of d_j, as described in Appendix.

A point k_i from X^j is then assigned the same cluster of $seed_j$ if the closeness between $seed_j$ and k_i is less than or equal to ε_j:

$$k_i \in \zeta_j \text{ , if } d_j(i) \le \varepsilon_j; \text{ otherwise } k_i \notin \zeta_j$$

6. PREDICTING VISUAL TARGET DISTINCTNESS

This section presents a computational visual distinctness measure computed from the image representational model based on visual patterns.

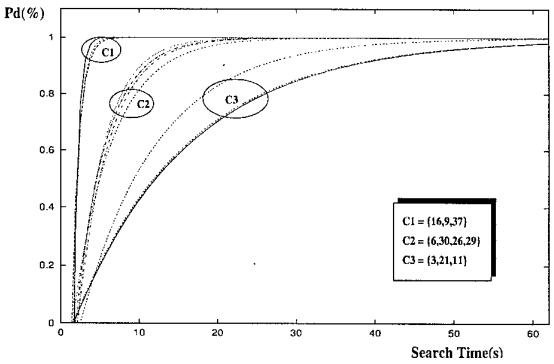


Figure 5: Cumulative distribution functions to the search times for the target scenes

The approach is as follows. First, a psychophysical experiment is performed in which observers estimate the visual distinctness of the target in each of 44 different test scenes (Section 6.2.). Second, a computational measure is defined and then applied to quantify the visual distinctness of the targets (Section 6.3.). Finally, an experiment is performed to investigate the relation between the computational distinctness measure and the visual target distinctness measured by human observers (Section 6.4.).

6.1. Images

The images used in this study are slides made during the DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test, that was designed and organized by NVESD (Night Vision & Electro-optic Sensors Directorate, Ft. Belvoir, VA, USA) and that was held in May and June 1995 in Fort Hunter Liggett, California, USA [15].These slides depict 44 different scenes.

Each scene represents a military vehicle in a complex rural background.The 9 different vehicles that are deployed as search targets are respectively a BMP-1, a BTR-70, an HMMVV-Scout, a HMMVV-Tow, an M1A1, an M3-Bradley, an M60, an M113, and a T72. The visibility of the targets varies throughout the entire stimulus set. This is mainly due to variations in the structure of the local background, the viewing distance, the luminance distribution over the target support (shadows), the orientation of the targets, and the degree of occlusion of the targets by vegetation.

The images used in the computational experiments are subsampled to 256x256 pixels. For each scene t, containing a target (vehicle), a corresponding empty scene e was created [6].The empty scene is everywhere equal to the target scene, except at the location of the target, where the target support is filled with the local background.This replacement is done by hand, using the rubber stamp tool in Photoshop 3.05.The result is judged by eye and is accepted if the variation in the background over the target support area does not appear to have an appreciable contrast with the natural variation in the local background.

In the experiment here reported (Section 6.4), the digital images were (see Figs. 6-9): (i) ten complex natural images containing a single target that correspond to the scenes 16, 9, 37, 6, 30, 26, 29, 3, 21, and 11, from the 44 slides made during the DISSTAF field test; and (ii) the corresponding empty images of the same rural backgrounds without target, that were created using the rubber stamp tool in Photoshop 3.05.

For each target image, Figs. 6-9 illustrate the simple thresholding of the visual pattern produced by the natural cluster of filters in *Active* that segregates the military vehicle

(target detector). Simple thresholding was applied to remove small response values which were present in the output of the target detector. In the same figures, it is also shown the visual pattern's thresholding produced by the target detector when it is applied on the respective image without target.
To produce the results shown in Figs. 6-9, the definition of integral feature used for the partitioning of the visual patterns in accord with a constraint of invariance
was as follows (Section 4.1):

- for the target scenes 30, 37, and 16, $T = (T_1, T_3, T_5)$;
- for 26 and 6, $T = (T_5)$;
- for 9, $T = (T_1, T_2, T_4)$;
- for 29, $T = (T_1, T_4, T_5)$;
- for 3, $T = (T_3)$;
- for 21, $T = (T_4, T_5)$; and
- for 11, $T = (T_1, T_2)$.

Section 6.4 analyzes how the best definition of integral feature for predicting visual target distinctness can be estimated on a dataset of example.

6.2. Psychophysical target distinctness

A psychophysical experiment was performed in which observers estimate the visual distinctness of the target. Search times and cumulative detection probabilities were measured for nine military targets in complex natural backgrounds.A total of 64 civilian observers, aged between 18 and 45 years, participate in the visual search experiment. The procedure of the search experiment is described in [6]. Search performance is usually expressed as the cumulative detection probability as function of time, and it can be approximated by [6] :

$$P_d(t)=\begin{cases}0 & , t<t_0\\ 1-\exp\{-(t-t_0)/\rho\} & , t\geq t_0\end{cases} \qquad (22)$$

where

- $P_d(t)$ is the fraction of correct detections at time t,
- t_0 is the minimum time required to response, and
- ρ is a time constant.

Fig. 5 shows the cumulative distribution functions corresponding to the search times measured for the target scenes used in the experiment here described. The overall difference between two of these functions can be measured by subtracting the area beneath their graphs. This operation corresponds to a Kolmogorov-Smirnov (K-S) test. To compare the relative distinctness of the targets in the different target scenes the curves are rank-ordered according to the area beneath their graphs. The resulting rank order for the target scenes is listed in the column with the header R_{P_d} in Table 1. These rank orders are adopted as the reference standard for the evaluation of the computational metric.
Targets that give rise to closely spaced cumulative detection curves which are similar in accordance with a K-S test, have similar visual distinctness. Fig. 5 shows that the target images in the experiment are clustered into a number of sets of targets with comparable visual distinctness: {16, 9, 37}, {6, 30, 26, 29}, and {3, 21, 11}. Consequently, rank order permutations of elements of the same cluster are not very significant, whereas rank order permutations of elements of different clusters are therefore significant.

6.3. Computational Target Distinctness

Let $C_n = \{\phi_{nj}\}$, with $n=1, 2, ..., N$, be the N natural clusters in *Active* produced by the RGFF model for the target image $t(x,y)$.

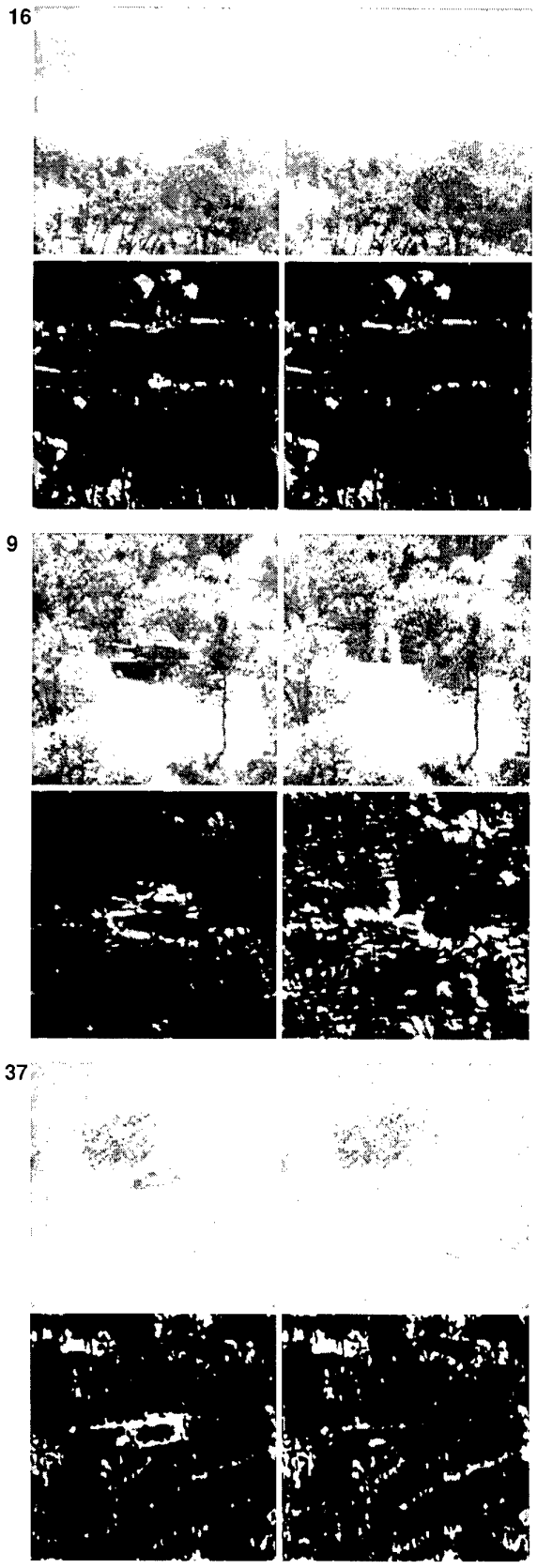


Figure 6: Target and empty images. Simple thresholding of the visual patterns produced by the target detetector on them.

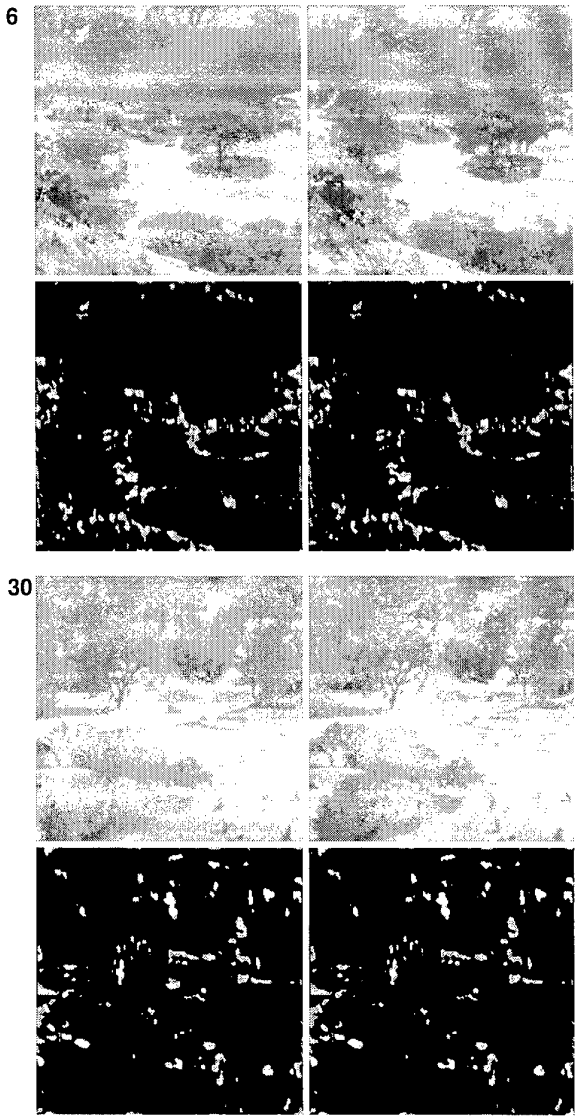


Figure 7: Target and empty scenes in the dataset, and the simple thresholding of the visual patterns produced by the target detector when it is applied on them. Simple thresholding was used to remove values which were present in the output of the target detector.

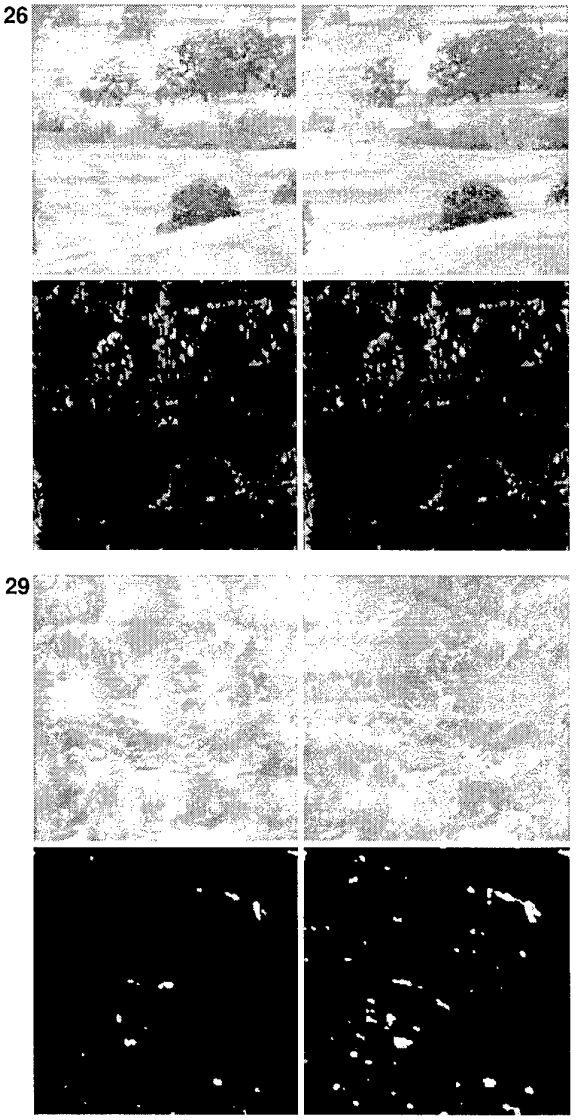


Figure 8: Target and empty scenes in the dataset, and the simple thresholding of the respective visual patterns by the target detector.

Let t_n represent the visual pattern segregated on the reference target image $t(x,y)$ by pooling the responses of filters in the natural cluster $C_n = \{\phi_{nj}\}$ as follows:

$$t_n = \left| \sum_j A_{n_j} \right| \tag{23}$$

where A_{nj}, denotes the original image $t(x,y)$ filtered through the logGabor ϕ_{nj} in C_n and passed through a non-linearity of the form:

$$tanh(z,\tau) = \frac{1-\exp\{-z\tau\}}{1+\exp\{-z\tau\}} \tag{24}$$

where τ is a gain term [16]. This nonlinearity enables the system to respond to local contrast over several log units of illumination changes.

Therefore $t_1, t_2, \ldots, t_N$ represent a decomposition of the reference target image t into the set of its most significant visual patterns.

In order to compensate for the effect of image-to-image variations on the overall image light level, contrast normalization of each visual pattern is realized by dividing t_n by the sum of all filtered responses in *Active*, plus a saturation constant σ:

$$\frac{t_n}{\sigma^2 + \sum_i |A_i|} \tag{25}$$

where A_i denotes the original image $t(x,y)$ filtered through the logGabor ϕ_i in *Active* and passed through a non-linearity as given in equation (24).

Similarly passing the corresponding empty image $e(x,y)$ through the filters associated with each cluster C_n produced by the model on the reference image $t(x,y)$, results in a decomposition of e in $e_1, e_2, \ldots, e_N$.

Let $d_{VP}(t_n, e_n)$ be the difference between the visual patterns t_n and e_n, computed via the β-norm between their statistical structure over those pixels which form "fixation points" on t_n [11]:

$$d_{VP}(t_n,e_n)=\frac{1}{Card\left[FP(t_n)\right]}\left(\sum_{(x,y)\in FP(t_n)}\left|D\left[T^{t_n}(x,y).T^{e_n}(x,y)\right]\right|^{\beta}\right)^{\frac{1}{\beta}} \quad (26)$$

with $FP(t_n)$ being the set of fixation points for t_n; and $D[T^{t_n}(x,y), T^{e_n}(x,y)]$ defining a normalized distance measure between the integral features $T^{t_n}(x,y)$ and $T^{e_n}(x,y)$ computed on t_n and e_n, respectively. The default value of the exponent β in Equation (26) is 3.

Based on a definition of "visual pattern" as congruence in T across frequency bands, the differences between segregated visual patterns, $D_n=d_{VP}(t_n,e_n)$, $n=1,2,\ldots, N$, determine the overall distinctness between the reference target image t and the corresponding empty image e by using a simple decision rule:

$$VP_T(t,e)=\frac{1}{N}\sum_{n=1}^{N} D_n \quad (27)$$

A schematic overview of the VP_T distinctness measure is given in Fig. 10.

6.4. Relation between the computational and psychophysical target distinctness estimates

All the possible definitions of T were considered by recombining any subset of the next separable features:

- the phase T_1,
- the local energy T_2,
- the standard deviation of the local energy T_3,
- the local contrast of the local energy T_4, and
- the entropy of the local energy T_5.

For each specific definition of integral feature, noted as T, the notion of congruence in T across frequency bands was used to decompose the images into its visual patterns. The VP_T measure was then applied to quantify the visual distinctness of the targets. The subjective ranking induced by the psychophysical target distinctness was the reference rank order.

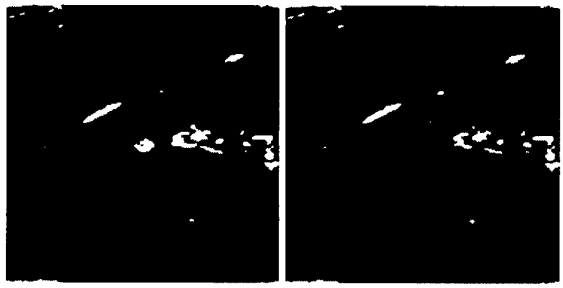
In order to study the efficacy of each definition T of integral feature for predicting target distinctness in a complex natural background, the fraction of correctly classified targets (with respect to the reference rank order) by the VP_T measure was computed on the dataset. Targets that give rise to closely spaced cumulative detection curves which are similar in accordance with a Kolmogorov-Smirnov test, have similar visual distinctness (Section 6.2.). Hence, the fraction of correct classification P_{CC} was defined as:

$$P_{CC}=\frac{Number\ of\ Correctly\ Classified\ Targets}{Number\ of\ Targets}$$

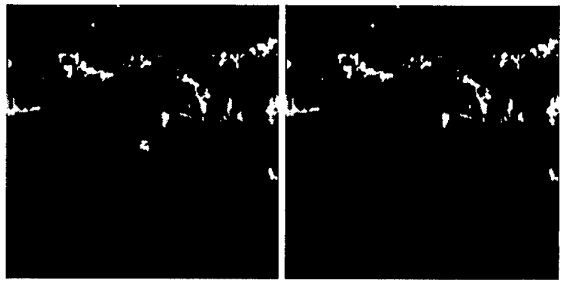
where rank order permutations of targets of the same cluster are insignificant (i.e., they are correctly classified by the metric), whereas rank order permutations of elements of different clusters are significant (the targets are then incorrectly classified).

The highest value of the fraction of correctly classified targets (P_{CC}=0.8) is obtained by the VP_T measure at $T=(T_1,T_2,T_3,T_4,T_5)$. Hence, the best definition of integral

3



21



11

Figure 9: Target and empty images. Thresholding of the visual patterns produced by the target detector.

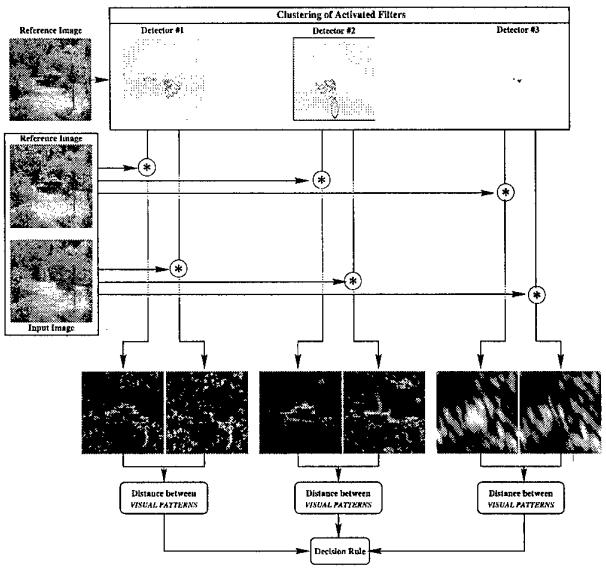


Figure 10: Schematic overview of the computational distinctness measure

feature for perceiving target distinctness on the dataset in this experiment, is $T = (T_1, T_2, T_3, T_4, T_5)$.

The comparative results of the *RMSE* metric and the VP_T measure based on the best definition of integral feature for predicting visual target distinctness are presented in Table 1. At the bottom of each of the columns is shown the respective fraction of correct classification. The reference rank order is listed in column 2.

The target distinctness values and the resulting rank order computed by the root mean square error (*RMSE*) metric are listed in column 3. The *RMSE* performs poorly, which is to be expected. Significant rank order permutations are displayed in boxes. The *RMSE* metric produces a rank order with five significant order reversals: targets 16, 26, 29, 3, and 11, are significantly out of order relative to the reference order induced by the psychophysical distinctness measure in column 2. The other targets have been attributed rank orders which do not differ significantly from the reference rank order. The *RMSE* yields a relatively low probability (P_{CC}=0.5). These results show that the *RMSE* metric appears not capable to rank order targets in the dataset with respect to their visual distinctness.

The target distinctness values and the resulting rank order computed by the $VP_{(T_1, T_2, T_3, T_4, T_5)}$ measure are listed in column 4. As noted above, this measure yields the highest probability (P_{CC}=0.8). This measure induces a rank order with two significant order reversals: targets 29 and 11 are ordered incorrectly. The other targets have been attributed rank orders which do not differ significantly from the reference rank order based on the psychophysical measure.

Summarizing, for the dataset in this experiment, the VP_T measure with $T=(T_1, T_2, T_3, T_4, T_5)$ appears to compute a visual target distinctness rank ordering that correlates with human observer performance.

7. CONCLUSION

Here a filtering technique was presented for the automatically learned partitioning of "visual patterns" in a digital image. Log-Gabor functions were adopted as an appropriate method to construct filters of arbitrary bandwidth. The novelty of our proposal lies in the definition of "visual patterns" as features which have the highest degree of alignment in statistical structure across different frequency bands. The interesting point is what kind of objects when imaged by cameras give rise to the visual patterns that the RGFF model segregates. They will be objects whose statistical structure across scales and orientations can be distinguished fairly well from the rest in a natural way. This limitation of the approach comes from the following assumption made in the clustering scheme of the Learning stage (Section[5]): the data set of activated filters has several separable clusters (e.g., elongated and non-piecewise linear separable groupings of arbitrary shape, dense and sparse natural clusters) and the membership is determined fairly well in a natural way by the data. The clarity of separation between clusters, as measured by a dissimilarity function, was the criterion by which they were derived. This assumption was needed to deal with several problems: (a) to overcome the lack of knowledge about the number and size of the clusters in the data, (b) to avoid the dependence of clustering on the initial cluster distribution, and (c) to find elongated and non-piecewise linear separable clusters, as well as to identify dense and sparse ones. In any case, the existence of natural clusters in the data is a very realistic assumption to many interesting applications. For example, because of the differences between the statistical structure across scales and orientations of targets and rural background in the application described in Section [6], the visual distinctness of a man-made object (a military vehicle) in a rural background can be determined in a natural way by the data.

Finally, a computational visual distinctness measure was presented that is computed from the image representational model based on visual patterns.It was applied to quantify the visual distinctness of targets in complex natural scenes. This measure that applies a simple decision rule to the distances between segregated visual patterns, was shown to correlate strongly with visual distinctness of targets in a dataset, as estimated by human observers.

Acknowledgments.

The authors thank Dr. Alexander Toet (TNO Human Factors Research Institute, The Netherlands) for providing us with image data, search times, and cumulative detection probabilities from search experiments made during the DISSTAF field test.

This research was sponsored by the Spanish Board for Science and Technology (CICYT) under grant TIC97-1150.

References

1. Field, D.J. "Scale-invariance and Self-similar `Wavelet' Transforms: an Analysis of Natural Scenes and Mammalian Visual Systems", in Wavelets, Fractals, and Fourier Transforms, Eds. M. Farge, J.C.K. Hunt, and J.C. Vassilicos, Clarendon Press, Oxford, pp. 151--193, 1993.
2. Morrone, M.C. and Burr, D. C. "Feature detection in human vision: A phase-dependent energy model." Proc. R. Soc. Lond. B, Vol. 235, pp. 221--245, 1988.
3. Field, D.J. "Relations between the statistics of natural images and the response properties of cortical cells", Journal of The Optical Society of America A, Vol. 4, No. 12, pp. 2379--2394, 1987.
4. Wertheimer, M. "Principles of perceptual organization," in Readings in perception, pp. 115--135, Van Nostrand, Princeton, NJ, 1958.
5. Lowe, D.G. "Three-dimensional object recognition from single two-dimensional images." Artificial Intelligence, Vol. 31, pp. 355--395, 1987.

6. Toet, A. "Computing visual target distinctness", TNO-report TM-97-A039, TNO Human Factors Research Institute, pp. 74, 1997.
7. Graham, Norma. Visual Pattern Analyzers. Oxford Psychology Series, No. 16, Oxford University Press, 1989.
8. Kovesi, P. "Image features from phase congruency", Technical Report 95/4, Department of Computer Science, The University of Western Australia, 1995.
9. Robbins, B. "The detection of 2D image features using local energy", D. Phil. thesis, Department of Computer Science, The University of Western Australia, 1996.
10. Fdez-Valdivia, J., Garcia, J.A., Martinez-Baena, J, and Fdez-Vidal, X.R. "The Selection of Natural Scales in 2D Images Using Adaptive Gabor Filtering", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, pp. 458--469, 1998.
11. Fdez-Vidal, X.R., Garcia, J.A., Fdez-Valdivia, J., and Rodriguez-Sanchez, Rosa. "The role of integral features for perceiving image discriminability", Pattern Recognition Letters, Vol. 18, pp. 733--740, 1997.
12. Treisman, A.M., and Gelade, G. "A feature-integration theory of attention." Cognitive Psychology, vol. 12, pp. 97--136, 1980.
13. Quick, R.F. "A vector-magnitude model of contrast detection". Kybernetik, 16, pp. 65--67, 1974.
14. Garcia, J.A., Fdez-Valdivia, J., Cortijo, F.J., Molina, R. "A dynamic approach for clustering data", Signal Processing, Vol. 44, pp. 181--196, 1995.
15. Toet, A., Bijl, P., Kooi, F.L., Valeton. J.M. "Image data set for testing search and detection models", TNO-report TM-97-A036, TNO Human Factors Research Institute, pp. 35, 1997.
16. Malik, J., and Perona, P. "Preattentive texture discrimination with early vision mechanisms", J. of Opt. Soc. Am. A, Vol. 7, No. 5, pp. 923--932, 1990.
17. Fdez-Valdivia, J., Garcia, J.A., and Garcia-Silvente, M. "An evaluation of the novel normalized-redundancy representation for planar curves", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 10, No. 7, pp. 769--789, 1996.
18. Witkin A. P. "Scale-Space Filtering". Proc. 8th Int. Joint. Conf. on Artificial Intelligence, Karlsruhe, West Germany, pp. 559--562, 1983.
19. Koenderink J.J., "The structure of images", Biological Cybernetics, Vol. 50, pp. 336--370, 1984.

APPENDIX

Let d_j'' be the second derivative of d_j computed as:

$$d_j''(i) = d_j(i) * \frac{d^2}{di^2} G_s(i)$$

with

$$G_s(i) = \frac{1}{\sqrt{2}s} \exp\left\{-\frac{i^2}{2s^2}\right\}$$

and where d_j is convolved with the second derivative of the Gaussian at scale s, noted as $\frac{d^2}{di^2}(i,(i))$, to both smooth and differentiate the function.

The zero crossings of d_j'' correspond to positions at which the dissimilarity d_j undergoes a significant increment in its value. To locate the zero crossings marking a rise in d_j due to inter-cluster differences, the unwanted detail from intra-cluster differences must be removed by smoothing. The question is: how much smoothing should be performed? The derivative should be processed at the scale that best describes the increments in d_j due to inter-cluster differences, while removing spurious increments due to intra-cluster differences.Each interesting structure in d_j comes from a significant rise in d_j due to inter-cluster differences, and the best scale for describing the structrure should be based on its intrinsic redundancy across scales as follows [17].

Because structures of interest exist as significant entities over a certain range of scales [18,19], one expects to find some redundancy across the different scales if there exist significant structures in d_j. That is, a significant structure should have a greater similarity represented at its natural scales (the levels of resolution at which the structure can be perceived in d_j).

Two smoothed versions of d_j at successive scales will be correlated to the extent their structures are similar at the respective scales. And we can determine the degree of similarity by correlating the smoothed versions of d_j at successive scales.

Let $d_j^{s(l)}$ with $l=1,\ldots,L$ be the dissimilarity d_j smoothed by Gaussian kernels at several levels of smoothing $s(l)$ ranging in value from 1 to $s(L)$ and increasing by a constant of 0.5 from one level to the next. Then the normalized redundancy measure, denoted as $A(s(l))$, between two smoothed versions $d_j^{s(l)}$ and $d_j^{s(l+1)}$, at successive scales $s(l)$ and $s(l+1)$ can be computed by cross-correlating $d_j^{s(l)}$ and $d_j^{s(l+1)}$ as follows:

$$A(s(l)) = \frac{\left\langle d_j^{s(l)}, d_j^{s(l+1)} \right\rangle}{\| d_j^{s(l)} \| \; \| d_j^{s(l+1)} \|}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the Hilbert space of measurable: square-integrable one-dimensional functions, and the norm (energy) of $d_j^{s(l)}$ is given by $\|d_j^{s(l)}\|^2$.

This function $A(s(l))$ returns a value measuring the relative redundancy between the respective smoothed versions at two consecutive scales. Given two smoothed versions, $d_j^{s(l)}$ and $d_j^{s(l+1)}$, at successive degrees of smoothing $s(l)$, $s(l+1)$ of a signal d_j, the value of the normalized function $A(s(l))$ at $s(l)$ is fairly small if any essential structure in $d_j^{s(l)}$has been removed from $d_j^{s(l+1)}$. Hence, each location $s(l)$ of local minima in $A(s(l))$ determines a significant scale for representing a structure of interest in d_j (i.e., a significant rise in d_j due to inter-cluster differences).

Consecuently, in order to locate the zero crossings of d_j'' marking a significant rise in d_j, the second derivative of d_j is then computed at the smallest scale from the set of locations $s(l)$ of local minima in $A(s(l))$. The derivative processed at the smallest significant scale, still describes the increments in d_j due to inter-cluster differences, while removing spurious increments due to intra-cluster differences.

TARGET DETECTION USING SALIENCY-BASED ATTENTION

Laurent Itti and Christof Koch
Computation and Neural Systems Program
California Institute of Technology
Mail-Code 139-74 - Pasadena, CA 91125 - U.S.A.
Email: {itti, koch}@klab.caltech.edu

1. SUMMARY

Most models of visual search, whether involving overt eye movements or covert shifts of attention, are based on the concept of a "saliency map", that is, an explicit two-dimensional map that encodes the saliency or conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location. We describe a detailed computer implementation of such a scheme, focusing on the problem of combining information across modalities, here orientation, intensity and color information, in a purely stimulus-driven manner. We have successfully applied this model to a wide range of target detection tasks, using synthetic and natural stimuli. Performance has however remained difficult to objectively evaluate on natural scenes, because no objective reference was available for comparison. We here present predicted search times for our model on the Search2 database of rural scenes containing a military vehicle. Overall, we found a poor correlation between human and model search times. Further analysis however revealed that in 3/4 of the images, the model appeared to detect the target faster than humans (for comparison, we calibrated the model's arbitrary internal time frame such that no more than 2-4 image locations were visited per second). It hence seems that this model, which had originally been designed not to find small, hidden military vehicles, but rather to find the few most obviously conspicuous objects in an image, performed as an efficient target detector on the Search2 dataset.

Keywords: Visual attention, saliency, preattentive, inhibition of return, model, winner-take-all, bottom-up, natural scene.

2. INTRODUCTION

Biological visual systems are faced with, on the one hand, the need to process massive amounts of incoming information (estimated at around 10^8 bits per second in the optic nerve of humans), and on the other hand, the requirement for nearly real-time capacity of reaction.

Surprisingly, instead of employing a purely parallel image analysis approach, primate vision systems appear to employ a serial computational strategy when inspecting complex visual scenes. Particular locations are selected based on their behavioral relevance or on local image cues. The identification of objects and the analysis of their spatial relationship usually involve either rapid, saccadic eye movements to bring the fovea onto the object, or covert shifts of attention. It consequently appears that the incredibly difficult problem of full-field image analysis and scene understanding is taken on by biological visual systems through a temporal serialization into smaller, localized analysis tasks.

Much evidence has accumulated in favor of a two-component framework for the control of where in a visual scene attention is focused to [1,2,3,4]: A bottom-up, fast, primitive mechanism that biases the observer towards selecting stimuli based on their "saliency" (most likely encoded in terms of center-surround mechanisms) and a second slower, top-down mechanism with variable selection criteria, which directs the "spotlight of attention" under cognitive, volitional control.

Koch and Ullman [5] introduced the idea of a saliency map to accomplish preattentive selection (see also the concept of a "master map" in [6]). This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics.

We here describe a computer implementation of a preattentive selection mechanism based on the architecture of the primate visual system. We address the thorny problem of how information from different modalities - in the case treated here from 42 maps encoding intensity, orientation and color in a center-surround fashion at a number of spatial scales - can be combined into a single saliency map. Our algorithm qualitatively reproduces human performance on a number of classical search experiments.

Vision algorithms frequently fail when confronted with realistic, cluttered images. We therefore studied the performance of our search algorithm using high-resolution (6144x4096 pixels) photographs containing images of military vehicles in a complex rural background (Search2 dataset). Our algorithm shows, on average, superior performance compared to human observers searching for the same targets, although our system does not yet include any top-down task-dependent tuning.

3. THE MODEL

The model has been presented in more details in [8] and is only briefly described here (Fig. 1).

Input is provided in the form of digitized color images. Different spatial scales are created using Gaussian pyramids [7], which consist of progressively low-pass filtering and subsampling the input image. Pyramids have a depth of 9 scales, providing horizontal and vertical image reduction factors ranging from 1:1 (scale 0; the original input image) to 1:256 (scale 8) in consecutive powers of two. Each feature is computed by center-surround operations akin to visual receptive fields, implemented as differences between a fine and a coarse scale: the center of the receptive field corresponds to a pixel at scale c={2, 3, 4} in the pyramid, and the surround to the corresponding pixel at scale s=c+d, with d={3, 4}, yielding six feature maps for each type of feature. The differences between two images at different scales are obtained by oversampling

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

the image at the coarser scale to the resolution of the image at the finer scale.

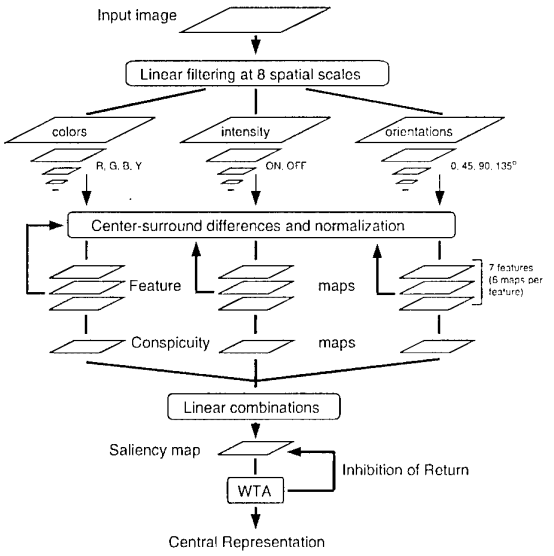


Figure 1: General architecture of the model. Low-level visual features are extracted in parallel from nine spatial scales, using a biological center-surround architecture. The resulting 42 feature maps are combined to yield three conspicuity maps for color, intensity and orientation. These, in turn, feed into a single saliency map, consisting of a 2D layer of integrate-and-fire neurons. A neural winner-take-all network shifts the focus of attention to the currently most salient image location. Feedback inhibition then transiently suppresses the currently attended location, causing the focus of attention to shift to the next most salient image location.

3.1. Extraction of early visual features

With r, g and b being the red, green and blue channels of the input image, an intensity image I is obtained as I=(r+g+b)/3. From I is created a Gaussian pyramid I(s), where s={0..8} is the scale. The r, g and b channels are normalized by I, at the locations where the intensity is at least 10% of its maximum, in order to decorrelate hue from intensity. Four broadly tuned color channels are created: R=r-(g+b)/2 for red, G=g-(r+b)/2 for green, B=b-(r+g)/2 for blue, and Y=(r+g)/2-|r-g|/2-b for yellow (negative values are set to zero). Four Gaussian pyramids R(s), G(s), B(s) and Y(s) are created from these color channels. From I, four orientation-selective pyramids are also created using Gabor filtering at 0, 45, 90 and 135 degrees.

Differences between a "center" fine scale c and a "surround" coarser scale s yield six feature maps for each of intensity contrast, red-green double opponency, blue-yellow double opponency, and the four orientations. A total of 42 feature maps is thus created, using six pairs of center-surround scales in seven types of features.

3.2. The saliency map

The task of the saliency map is to compute a scalar quantity representing the salience at every location in the visual field, and to guide the subsequent selection of attended locations. The feature maps provide the input to the saliency map, which is modeled as a neural network receiving its input at scale 4.

3.2.1. Fusion of information

One difficulty in combining different feature maps is that they represent a priori not comparable modalities, with different dynamic ranges and extraction mechanisms. Also, because a total of 42 maps are combined, salient objects appearing strongly in only a few maps risk to be masked by noise or less salient objects present in a larger number of maps.

Previously, we have shown that the simplest feature combination scheme - to normalize each feature map to a fixed dynamic range, and then sum all maps - yields very poor detection performance for salient targets in complex natural scenes [9]. One possible way to improve performance is to learn linear map combination weights, by providing the system with examples of targets to be detected. While performance improves greatly, this method presents the disadvantage of yielding different specialized models (that is, sets of map weights) for each target detection task studied [9].

When no top-down supervision is available, we propose a simple normalization scheme, consisting of globally promoting those feature maps in which a small number of strong peaks of activity (conspicuous locations) is present, while globally suppressing feature maps which contain comparable peak responses at numerous locations over the visual scene. This "within-feature competitive" scheme coarsely ressembles non-classical inhibitory interactions which have been observed electrophysiologically [10].

The specific implementation of these interactions in our model has been described elsewhere [9] and can be summarized as follows (Fig. 2): Each feature map is first normalized to a fixed dynamic range (between 0 and 1), in order to eliminate feature-dependent amplitude differences due to different feature extraction mechanisms. Each feature map is then iteratively convolved by a large 2-D Derivative-of-Gaussians (DoG) filter. The DoG filter, a section of which is shown in Fig. 2, yields strong local excitation at each visual location, which is counteracted by broad inhibition from neighboring locations. At each iteration, a given feature map receives input from the preattentive feature extraction stages described above, to which results of the convolution by the DoG are added. All negative values are then rectified to zero, thus making the iterative process highly non-linear. This procedure is repeated for 10 iterations.

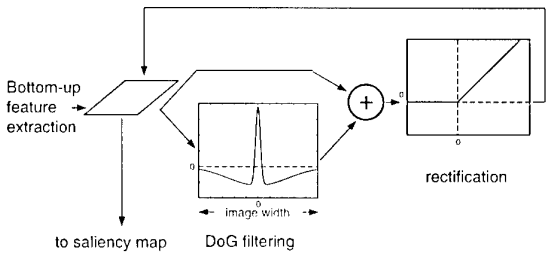


Figure 2: Illustration of the spatial competition for salience implemented within each of the 42 feature maps. Each map receives input from the linear filtering and center-surround stages. At each step of the process, the convolution of the map by a large Difference-of-Gaussians (DoG) kernel is added to the current contents of the map. This additional input coarsely models short-range excitatory processes and long-range inhibitory interactions between neighboring visual locations. The map is half-wave rectified, such that negative values are eliminated, hence making the iterative process non-linear. Ten iterations of the process are carried out before the output of each feature map is used in building the saliency map.

The choice of the number of iterations is somewhat arbitrary: In the limit of an infinite number of iterations, any non-empty map will converge towards a single peak, hence constituting only a poor representation of the scene. With very few iterations however, spatial competition is very weak and inefficient. Two examples showing the time evolution of this process are shown in Fig. 3, and illustrate that using of the order of 10 iterations yields adequate distinction between the two example images shown. As expected, feature maps with initially numerous peaks of similar amplitude are suppressed by the interactions, while maps with one or a few initially stronger peaks become enhanced. It is interesting to note that this within-feature spatial competition scheme resembles a "winner-take-all" network with localized inhibitory spread, which allows for a sparse distribution of winners across the visual scene.

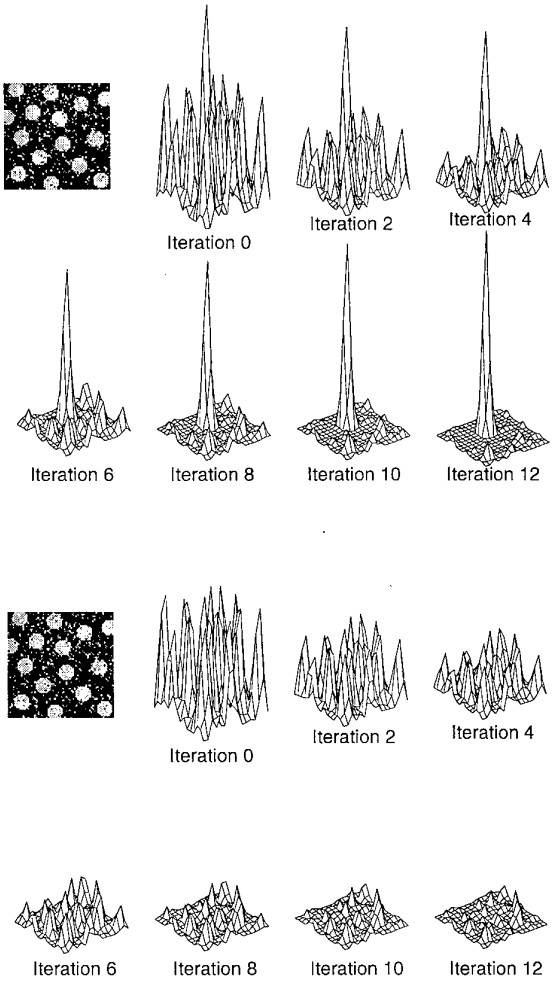


Figure 3: Example of operation of the long-range iterative competition for salience. When one (or a few) locations elicit stronger responses, they inhibit more the other locations than they are inhibited by these locations; the net result after a few iterations is an enhancement of the initially stronger location(s), and a suppression of the weaker locations. When no location is clearly stronger, all locations send and receive approximately the same amount of inhibition; the net result in this case is that all locations progressively become inhibited, and the map is globally suppressed.

After normalization, the feature maps for intensity, color, and orientation are summed across scales into three separate "conspicuity maps", one for intensity, one for color and one for orientation (Fig. 1).

Each conspicuity map is then subjected to another 10 iterations of the iterative normalization process. The motivation for the creation of three separate channels and their individual normalization is the hypothesis that similar features compete strongly for salience, while different modalities contribute independently to the saliency map. Although we are not aware of any supporting experimental evidence for this hypothesis, this additional step has the computational advantage of further enforcing that only a spatially sparse distribution of strong activity peaks is present within each visual feature, before combination of all three features into the scalar saliency map.

3.2.2. Internal Dynamics And Trajectory Generation

By definition, at any given time, the maximum of the saliency map's neural activity is at the most salient image location, to which the focus of attention (FOA) should be directed. This maximum is detected by a winner-take-all (WTA) network inspired from biological architectures [5]. The WTA is a 2D layer of integrate-and-fire neurons with a much faster time constant than those in the saliency map, and with strong global inhibition reliably activated by any neuron in the layer. In order to create dynamical shifts of the FOA, rather than permanently attending to the initially most salient location, it is necessary to transiently inhibit, in the saliency map, a spatial neighborhood of the currently attended location. This also prevents the FOA from immediately coming back to a strong, previously attended location. Such an "inhibition of return" mechanism has been demonstrated in humans [11]. Therefore, when a winner is detected by the WTA network, it triggers three mechanisms (Fig. 4):

1) The FOA is shifted so that its center is at the location of the winner neuron;

2) The global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons;

3) Inhibitory conductances are transiently activated in the saliency map, in an area corresponding to the size and new location of the FOA. In order to slightly bias the model to next jump to salient locations spatially close to the currently attended location, small excitatory conductances are also transiently activated in a near surround of the FOA in the saliency map ("proximity preference" rule proposed by Koch and Ullman [5]).

Since we do not model any top-down mechanism, the FOA is simply represented by a disk whose radius is fixed to one twelvth of the smaller of the input image width or height. The time constants, conductances, and firing thresholds of the simulated neurons are chosen so that the FOA jumps from one salient location to the next in approximately 30-70ms (simulated time), and so that an attended area is inhibited for approximately 500-900ms, as it has been observed psychophysically [11]. The difference in the relative magnitude of these delays proved sufficient to ensure thorough scanning of the image by the FOA and prevent cycling through a limited number of locations.

Fig. 4 demonstrates the interacting time courses of two neurons in the saliency map and the WTA network, for a very simple stimulus consisting of one weaker and one stronger pixels in an otherwise empty map.

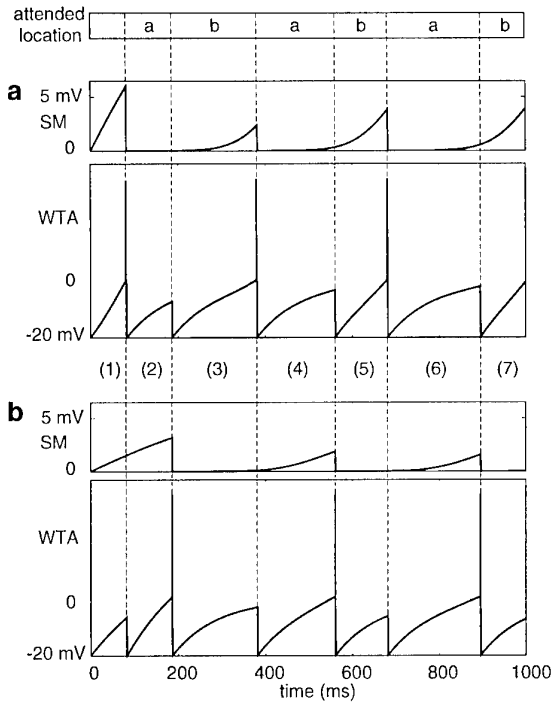


Figure 4: Dynamical evolution of the potential of some simulated neurons in the saliency map (SM) and in the winner-take-all (WTA) networks. The input contains one salient location (a), and another input of half the saliency (b); the potentials of the corresponding neurons in the SM and WTA are shown as a function of time. During period (1), the potential of both SM neurons (a) and (b) increases as a result of the input. The potential in the WTA neurons, which receive inputs from the corresponding SM neurons but have much faster time constants, increases faster. The WTA neurons evolve independently of each other as long as they are not firing. At about 80ms, WTA neuron (a) reaches threshold and fires. A cascade of events follows: First, the focus of attention is shifted to (a); second, both WTA neurons are reset; third, inhibition-of-return (IOR) is triggered, and inhibits SM neuron (a) with a strength proportional to that neuron's potential (i.e., more salient locations receive more IOR, so that all attended locations will recover from IOR in approximately the same time). In period (2), the potential of WTA neuron (a) rises at a much slower rate, because SM neuron (a) is strongly inhibited by IOR. WTA neuron (b) hence reaches threshold first. (3)-(7): In this example with only two active locations, the system alternatively attends to (a) and (b). Note how the IOR decays over time, allowing for each location to be attended several times. Also note how the amount of IOR is proportional to the SM potential when IOR is triggered (e.g., SM neuron (a) receives more IOR at the end of period (1) than at the end of period (3)). Finally, note how the SM neurons do not have an opportunity to reach threshold (at 20 mV) and to fire (their threshold is ignored in the model). Since our input images are noisy, we did not explicitly incorporate noise into the neurons' dynamics.

4. RESULTS

4.1. General performance

We tested our model on a wide variety of real images, ranging from natural outdoor scenes to artistic paintings. All images were in color, contained significant amounts of noise, strong local variations in illumination, shadows and reflections, large numbers of "objects" often partially occluded, and strong textures. Most of these images can be interactively examined on the World-Wide-Web, at:
http://www.klab.caltech.edu/~itti/attention/

Overall, the results indicate that the system scans the image in an order which makes functional sense in most behavioral situations.

It should be noted however that it is not straightforward to establish objective criteria for the performance of the system with such images. Unfortunately, nearly all quantitative psychophysical data on attentional control are based on synthetic stimuli. In addition, although the scan paths of overt attention (eye movements) have been extensively studied [12], it is unclear to what extent the precise trajectories followed by the attentional spotlight are similar to the motion of covert attention. Most probably, the requirements and limitations (e.g., spatial and temporal resolutions) of the two systems are related but not identical [13].

Although our model is mostly concerned with shifts of covert attention, and ignores all of the mechanistic details of eye movements, we attempt below a quantitative comparison between human and model target search times in complex natural scenes, using the Search2 database of images containing military vehicles hidden in a rural environment.

4.2. Search2 results

We propose a difficult test of the model using the Search2 dataset, in which target detection is evaluated using a database of complex natural images, each containing a military vehicle (the "target"). Contrary to our previous study with a simplified version of the model [8], which used low-resolution image databases with relatively large targets (typically about 1/10th the width of the visual scene), this study uses very-high resolution images (6144x4096 pixels), in which targets appear very small (typically 1/100th the width of the image). In addition, in the present study, search time is compared between the model's predictions and the average measured search times from 62 normal human observers [14].

4.2.1. Experimental setup

The 44 original photographs were taken during a DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test in Fort Hunter Liggett, California, and were provided to us, along with all human data, by the TNO Human Factors Research Institute in the Netherlands [14]. The field of view for each image is 6.9x4.6 deg. Each scene contained one of nine possible military vehicles, at a distance ranging from 860 to 5822 meters from the observer. Each slide was digitized at 6144x4096 pixels resolution. Sixty two human observers aged between 18 and 45 years and with visual acuity better than 1.25 arcmin^-1 participated to the experiment (about half were women and half men).

Subjects were first presented with 3 close-up views of each of the 9 possible target vehicles, followed by a test run of 10 trials. A Latin square design [14] was then used for the randomized presentation of the images. The slides were projected such that they subtended 65x46 deg visual angle to the observers (corresponding to a linear magnification by about a factor 10 compared to the original scenery). During each trial, observers pressed a button as soon as they had detected the target, and subsequently indicated at which location on a 10x10 projected grid they had found the target. Further details on these experiments can be found in [14].

The model was presented with each image at full resolution. Contrary to the human experiment, no close-ups or test trials were presented to the model. The generic form of the model described above was used, without any specific parameter adjustment for this experiment. Simulations for up to 10,000 ms. of simulated time (about 200-400 attentional shifts) were done on a Digital Equipment Alpha 500 workstation. With these high-resolution images, the model comprised about 300 million simulated neurons. Each image was processed in about 15 minutes with a peak memory usage of 484 megabytes (for comparison, a 640x480 scene was typically processed in 10 seconds, and processing time approximately scaled linearly with the number of pixels). The focus of attention (FOA) was represented by a disk of radius 340 pixels (Figs. 5, 6, 7). Full coverage of the image by the FOA would hence require 123 shifts (with overlap); a random search would thus be expected to find the target after 61.5 shifts on average. The target was considered detected when the focus of attention intersected a binary mask representing the outline of the target, which was provided with the images. Three examples of scenes and model trajectories are presented in Figs. 5, 6, and 7. In the one image, the target was immediately found by the model, in another, a serial search was necessary before the target could be found, and in the last, the model failed to find the target.

4.2.2. Simulation results

The model immediately found the target (first attended location) in seven of the 44 images. It quickly found the target (fewer than 20 shifts) in another 23 images. It found the target after more than 20 shifts in 11 images, and failed to find the target in 3 images. Overall, the model consequently performed surprisingly well, with a number of attentional shifts far below the expected 61.5 shifts of a random search in all but 6 images. In these 6 images, the target was extremely small (and hence not conspicuous at all), and the model cycled through a number of more salient locations.

4.2.3. Tentative comparison to human data

The following analysis was performed to generate the plot presented in Fig. 8: First, a few outlier images were discarded, when either the model did not find the target within 2000ms of simulated time (about 40-80 shifts; 6 images), or when half or more of the humans failed to find the target (3 images), for a total of 8 discarded images. An average of 40ms per model shift was then derived from the simulations, and an average of 3 overt shifts per second was assumed for humans, hence allowing us to scale the model's simulated time to real time. An additional 1.5 second was then added to the model time to account for human motor response time. With such calibration, the fastest reaction times for both model and humans were approximately 2 seconds, and the slowest approximately 15 seconds, for the 36 images analyzed.

The results plotted in Fig. 8 overall show a poor correlation between human and model search times. Surprisingly however, the model appeared to find the target faster than humans in 3/4 of the images (points below the diagonal), despite the rather conservative scaling factors used to compare model to human time. In order to make the model faster than humans in no more than half of the images, one would have to assume that humans shifted their gaze not faster than twice per second, which seems unrealistically slow under the circumstances of a speeded search task on a stationary, non-masked scene. Even if eye movements were that slow, most probably would humans still shift covert attention at a much faster rate between two overt fixations.

4.2.4. Comparison to spatial frequency content models

In our previous studies with this model, we have shown that the within-feature long-range interactions are one of the key aspects of the model. In order to illustrate this point, we can compute a simple measure of local spatial fequency content (SFC) at each location in the input image, and compare this measure to our saliency map.

It could indeed be argued that the preattentive, massively parallel feature extraction stages in our model constitute a simple set of spatially and chromatically bandpass filters. A possibly much simpler measure of "saliency" could hence be based on a more direct measure of power or of amplitude in different spatial and chromatic frequency bands. Such simpler measure has been supported by human studies, in which local spatial frequency content (measured by Haar wavelet transform) was higher at the points of fixations during free viewing than on average, over the entire visual scene (see [8] for details).

We illustrate in Fig. 9, with one representative example image, that our measure of saliency actually differs greatly from a simple measure of SFC. The SFC was computed as shown previously [8], by taking the average amplitude of non-negligible FFT coefficients computed for the luminance channel as well as the red, green, blue and yellow channels.

While the SFC measure shows strong responses at numerous locations, e.g., at all locations with sharp edges, the saliency map contains a much sparser representation of the scene, where only locally unique such regions are preserved.

5. DISCUSSION

We have demonstrated that a relatively simple processing scheme, based on some of the key organizational principles of pre-attentive early visual cortical architectures (center-surround receptive fields, non-classical within-feature inhibition, multiple maps) in conjunction with a single saliency map performs remarkably well at detecting salient targets in cluttered natural and artificial scenes.

Key properties of our model, in particular its usage of inhibition-of-return and the explicit coding of saliency independent of feature dimensions, as well as its behavior on some classical search tasks, are in good qualitative agreement with the human psychophysical literature.

Using reasonable scaling of model to human time, we found that the model appeared to find the target faster than humans in 75% of the 36 images studied. One paradoxical explanation for this superior performance might be that top-down influences play a significant role in the deployment of attention in natural scenes. Top-down cues in humans might indeed bias the attentional shifts, according to the progressively constructed mental representation of the entire scene, in inappropriate ways. Our model lacks any high-level knowledge of the world and operates in a purely bottom-up manner.

This does suggest that for certain (possibly limited) scenarios, such high-level knowledge might interfere with optimal performance. For instance, human observers are frequently tempted to follow roads or other structures, or may consciously decide to thoroughly examine the surroundings of salient buildings that have popped-out, while the vehicle might be in the middle of a field or in a forest.

Although our model was not originally designed to detect military vehicles, our results also suggest that these vehicles where fairly "salient", according to the measure of saliency implemented in the model. This is also surprising, since one would expect such vehicles to be designed **not** to be salient.

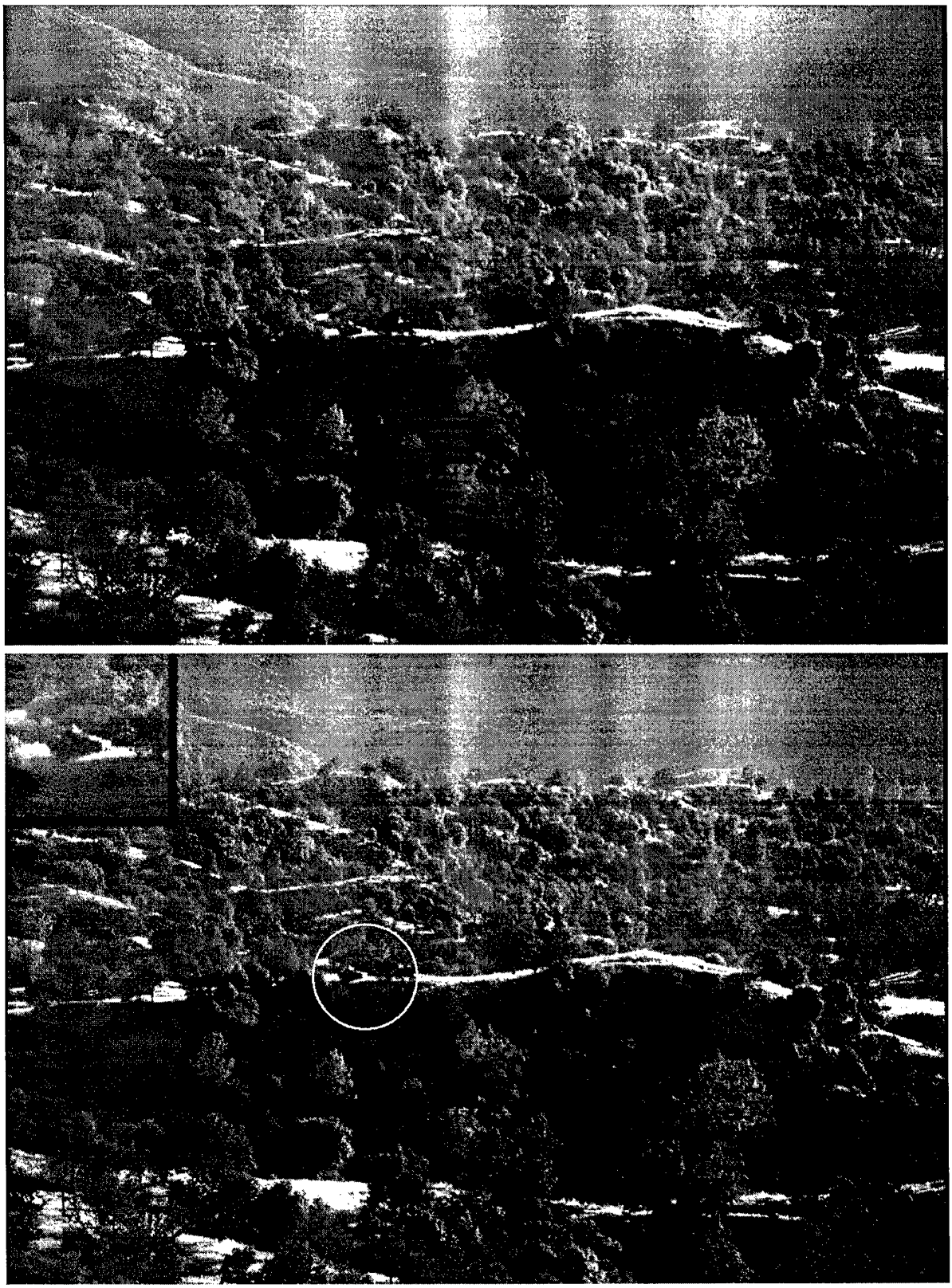


Figure 5: Example of image from the Search2 dataset (image 0018). The algorithm operated on the 24-bit color image. Top: original image; humans found the target in 2.8 sec on average. Bottom: model prediction: the target was the first attended location. After scaling of model time such that two to four attentional shifts occurred each second on average, and addition of 1.5 sec to account for latency in human motor response, the model found the target in 2.2 sec.

Figure 6: A more difficult example of image from the Search2 dataset (image 0019). Top: original image; humans found the target in 12.3 sec on average. Bottom: model prediction; because of its low contrast to the background, the target had lower saliency than several other objects in the image, such as buildings. The model hence initiated a serial search and found the target as the 10th attended location, after 4.9 sec (using the same time scaling as in the previous figure).

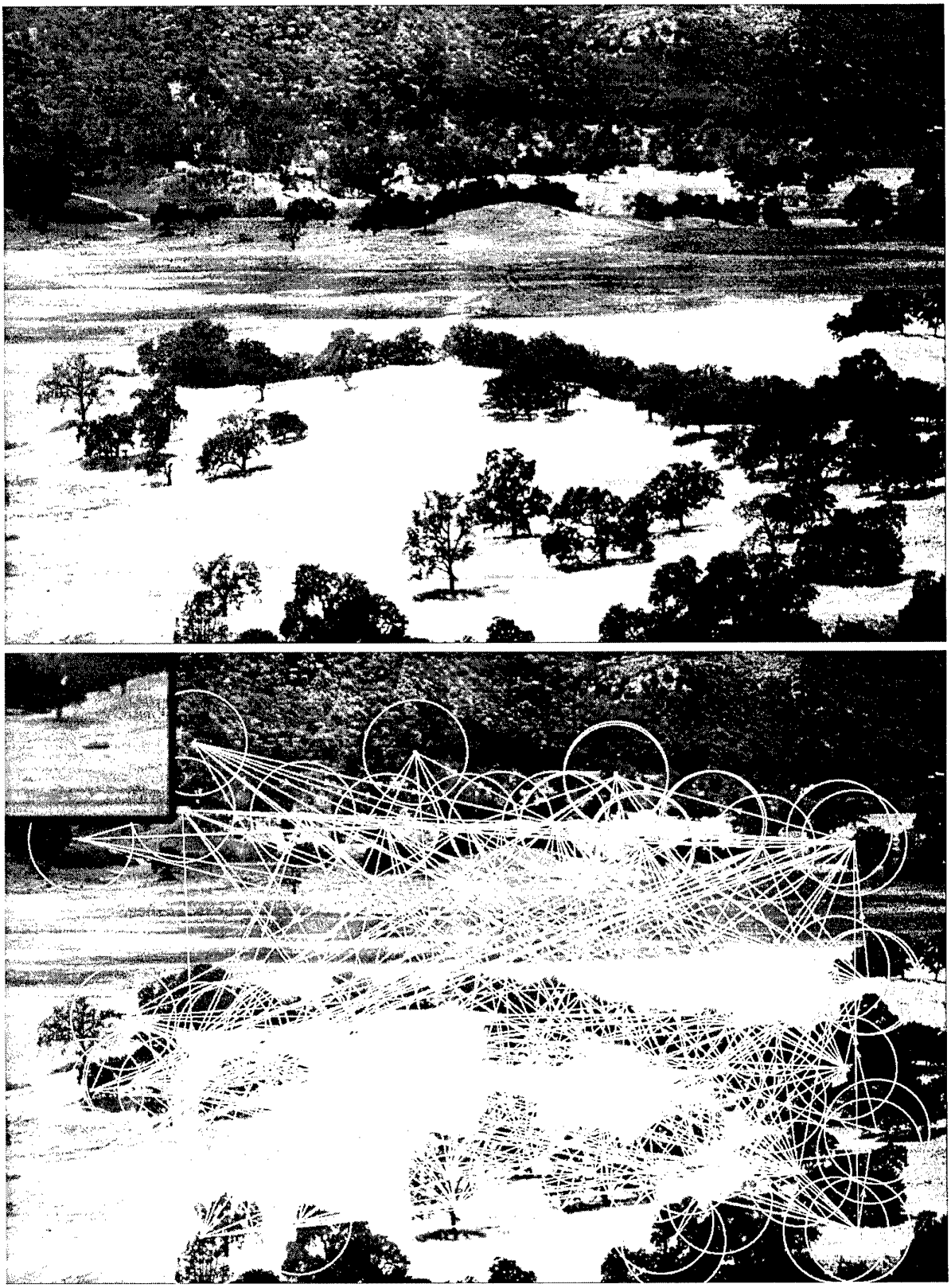


Figure 7: Example of image from the Search2 dataset (image 0024) in which the model did not find the target. Top: original image; humans found the target in 8.0 sec on average. Bottom: model prediction; the model failed to find the target, whose location is indicated by the white arrow. Inspection of the feature maps revealed that the target yielded responses in the different feature dimensions which are very similar to other parts of the image (foliage and trees). The target was hence not considered salient at all.

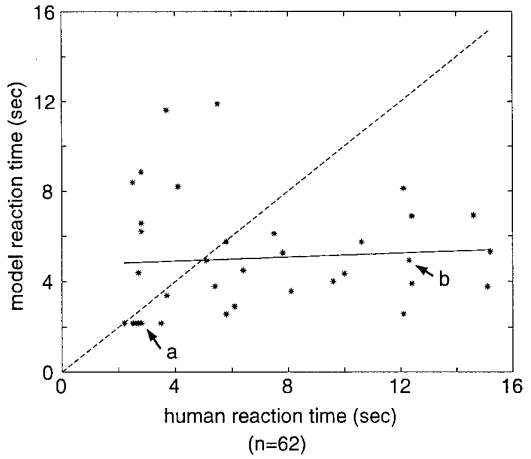


Figure 8. Mean reaction time to detect the target for 62 human observers and for our deterministic algorithm. Eight of the 44 original images are not included, in which either the model or the humans failed to reliably find the target. For the 36 images studied, and using the same scaling of model time as in the previous two figures, the model was faster than humans in 75% of the images. In order to bring this performance down to 50% (equal performance for humans and model), one would have to assume that no more than two visual locations can be visited each second. Arrow (a) indicates the "pop-out" example of Fig. 5, and arrow (b) the more difficult example presented in Fig. 6.

Looking at the details of individual feature maps, we realized that in most cases of quick detection of the target by the model, the vehicle was salient due to a strong, spatially isolated peak in the intensity or orientation channels. Such peak usually corresponded to the location of a specular reflection of sunlight onto the vehicle. Specular reflections were very rare at other locations in the images, and hence were determined to pop-out by the model. Because these reflections were often associated with locally rich SFC, and because many other locations also showed rich SFC, the SFC map could not detect them as reliably. Because these regions were spatially unique in one type of feature, they however popped-out for our model. Our model would hence have shown much poorer performance if the vehicles had not been so well polished.

6. CONCLUSION

In conclusion, our model yielded respectable results on the Search2 dataset, especially considering the fact that no particular adjustment was made to the model's parameters in order to optimize its target detection performance.

One important issue which needs to be addressed however is that of the poor correlation between model and human search times. We hypothesized in this study that top-down, volitional attentional bias might actually have hurt humans with this particular dataset, because trying to understand the scene and to willfully follow its structure was of no help in finding the target. A verification of this hypothesis should be possible once the scanpaths of human fixations during the search become available for the Search2 dataset.

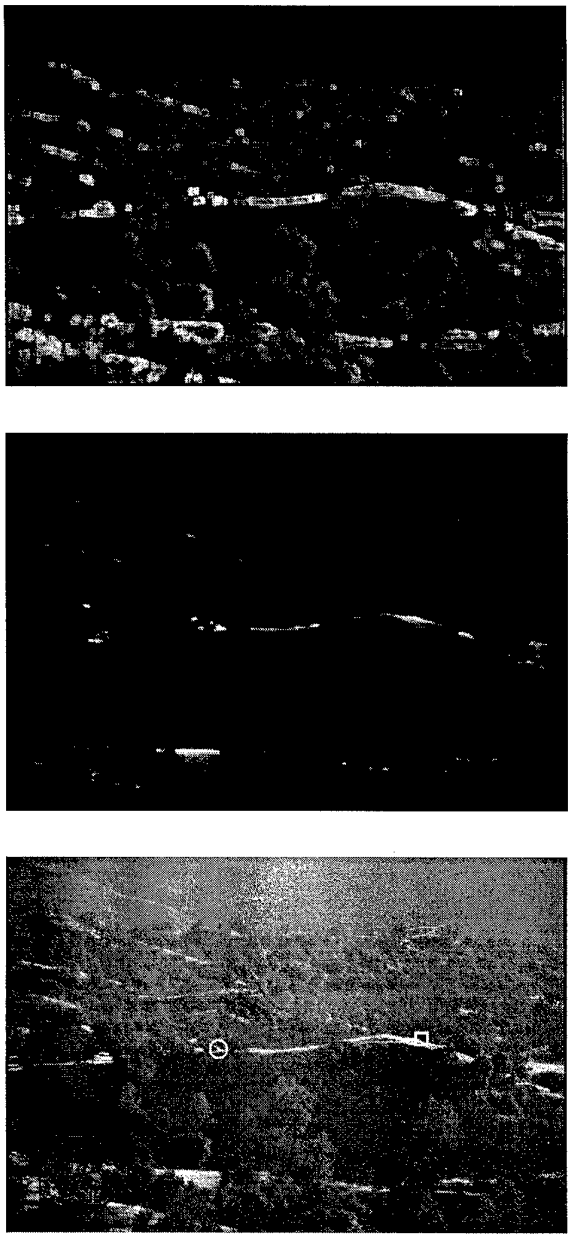


Figure 9: Comparison of SFC and saliency maps for image 0018 (shown in Fig. 5). Top: the SFC map shows strong response at all locations which have "rich" local textures; that is almost everywhere in this image. Middle: The within-feature, spatial competition for salience however demonstrates efficient reduction of information by eliminating large areas of similar textures. Bottom: The maximum of the saliency map (circle) is at the target, which appeared as a very strong isolated object in a few intensity maps because of the specular reflection on the vehicle. The maximum of the SFC map is at another location on the road.

7. ACKNOWLEDGMENTS

We thank Dr. A. Toet from TNO-HFRI for providing us with the search2 dataset and all human data. This work was supported by NSF (Caltech ERC), HIMH, ONR and NATO.

8. REFERENCES

1. James, W., *The Principles of Psychology*. Harvard Univ Press, Cambridge, MA, USA, 1890.
2. Treisman, A.M., Gelade, G., "A feature-integration theory of attention", *Cognit Psychol* 12, pp. 97-136, 1980.
3. Bergen, J.R., Julesz, B., "Parallel versus serial processing in rapid pattern discrimination", *Nature*, 303, pp. 696-8, 1983.
4. Barun, J., Julesz, B., "Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept Psychophys*, 60, pp. 1-23, 1998
5. Koch, C., Ullman, S., "Shifts in selective visual attention: towards the underlying neural circuitery". *Hum Neurobiol*, 4, pp. 219-27, 1985.
6. Treisman, A., "Features and objects: the fourteenth Bartlett memorial lecture", *Q J Exp Psychol [A]*, 40, pp. 201-37, 1998.
7. Burt, P., Adelson, E., "The laplacian pyramid as compact image code", *IEEE Trans on Comm*, 31, pp. 532-40, 1983.
8. Itti, L., Koch, C., Niebur, E., "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans Patt Anal Mach Intell*, 20, pp. 1254-9, 1998.
9. Itti, L., Koch, C., "A comparison of feature combination strategies for saliency-based visual attention", in *SPIE Human Vision and Electronic Imaging IV*, San Jose, CA, 1999 (in press).
10. Sillito, A.M., Grieve, K.L., Jones, H.E., Cudeiro, J., Davis, J., "Visual cortical mechanisms detecting focal orientation discontinuities", *Nature*, 378, pp. 492-6, 1995.
11. Posner, M.I., Cohen, Y., Rafal, R.D., "Neural systems control of spatial orienting", *Philos Trans R Soc Lond B Biol Sci*, 298, pp. 187-98, 1982.
12. Yarbus, A.L., *Eye movements and vision*. Plenum Press, New York, USA, 1967.
13. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K, Lai, Y.H., Davis, N., Nuflo, F. "Modeling visual-attention via selective tuning", *Artif Intell*, 78, pp.507-45, 1995.
14. Toet, A., Bijl, P., Kooi, F.L., Valenton, J.M., *A high-resolution image dataset for testing search and detection models* (Report TNO-TM-98-A020), TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.

APPLYING THE LAW OF COMPARATIVE JUDGEMENT TO TARGET SIGNATURE EVALUATION

James R. McManamey
U.S. Army Communications and Electronics Command
Night Vision and Electronic Sensors Directorate
10221 Burbeck Road, Suite 430, Building 305
Fort Belvoir, Virginia 22060-5806
E-mail: jmcmanam@nvl.army.mil

1. SUMMARY

The Law of Comparative Judgement (LCJ) is a psychophysical tool that can be used to scale complex phenomena that lack easily identified physical parameters. Target signatures represent such phenomena. In a demonstration exercise, a "search difficulty" value was found using the LCJ. These LCJ scale values were compared to search times and probabilities of detection from a search experiment run in the Netherlands. The scale values were not linearly related to search time and probability of detection, but correlated very well with the logarithm of mean search time ($r = 0.936$) and the cube of the number of correct responses ($r = 0.954$). A chi-squared goodness-of-fit test gave 94.6% confidence in the fit of the LCJ scale to the experimental data. While the LCJ results in a scale with no natural zero point and arbitrary units, this tool can be used to construct a standard scale. This paper illustrates how a standard clutter scale might be constructed using the LCJ. The LCJ could be a valuable tool in target signature evaluation either when used in conjunction with scaling equations that permit conversion to familiar quantities such as mean search time and probability of detection, by providing relative "search difficulty" values, or by making possible a psychophysically meaningful clutter scale.

Keywords: Law of Comparative Judgement, search difficulty, clutter, psychophysical methods, scaling methods, paired comparison, signature evaluation.

2. INTRODUCTION

Today, there are many quantities that engineers and scientists want to measure in perceptually meaningful ways. For example, designers of military man-in-the-loop search and target acquisition systems, as well as engineers working on military signature suppression systems, want measures of effectiveness that are psychophysically meaningful, repeatable, and correlate well with field performance. Such measures of effectiveness have frequently been surprisingly elusive. Target detectability and signature levels may seem like concrete, physically measurable quantities, but in truth they have much in common with such abstract concepts as beauty. Figure 1 shows a near-infrared scene. The upper image shows a tank profile that has been inserted into the scene. In the lower image, the tank is not visible at all. It is "perfectly camouflaged." However, most signature evaluation models and virtually all of the most widely used sensor models would say that the two tanks have exactly the same signature. This is because the only difference between these two target signatures is that the image pixels have been moved around. Averaged over the target, the histogram, contrast, variance, third-, fourth-, and fifth-moments are all the same. Only measures of effectiveness that can distinguish between the relatively large "blobs" in the lower image and the "salt-and-pepper" noise in the upper image can distinguish between the two tanks. Only a model that can determine that the tank in the lower image has the same "texture" as the background and that the edges are perfectly "blended" with the background while, at the same time, determining that these things are not true of the tank in the upper image, can accurately predict that a person will detect the target in the top picture and fail to detect the target in the lower one.

Investigators around the world are trying to develop models that can make such distinctions. Many of these models, a type called "computational vision models," attempt to mimic various processes that are believed to take place in the human eye-brain system. This has been a daunting task, and none of the computational vision models can really be considered complete, calibrated, and fully validated, although some of these models are validated for specific applications.

While we don't yet have models that can accurately and reliably predict detection probabilities throughout the range represented by the two images in figure 1, there are reliable scaling methods that can help to provide the correct signature level figures-of-merit in a wide variety of situations, including those depicted in this illustration. These scaling methods can provide the psychophysical values with which modeled quantities must correlate. One such method is the Law of Comparative Judgement (LCJ). The LCJ permits us to assign a one-dimensional scale to complex phenomena such as target signature levels even though they may lack an easily identified set of physical attributes and may frequently be a matter of opinion.

3. THE LAW OF COMPARATIVE JUDGEMENT

Between 1925 and 1932, Louis Thurstone published 24 articles and a book on how to construct good measurement scales. Today the name Thurstone is synonymous with scaling methods that result in equal-appearing intervals. One of his contributions to the field of psychology is the law of comparative judgement (LCJ).

In the beginning, the LCJ was a psychophysical tool for determining discrimination thresholds and psychological equivalents of physically measurable stimuli. For example, a subject could be presented with a tone of a particular pitch, loudness, and duration, followed by a second tone of the same pitch and duration but not the same loudness. The subject could then be asked whether the second tone was louder or softer than the first one. In this way, investigators could find out how sound pressure translates into perceived sensations. However, the LCJ provides only indirect scaling. As direct means were devised for measuring the same phenomena, psychophysicists turned to these direct methods and the role of the LCJ was gradually eroded. However, abstract sensations (attitudes, opinions, and aesthetic values) provided no physically measurable qualities. Finally, the LCJ came to be primarily a means of characterizing abstract stimuli[1].

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

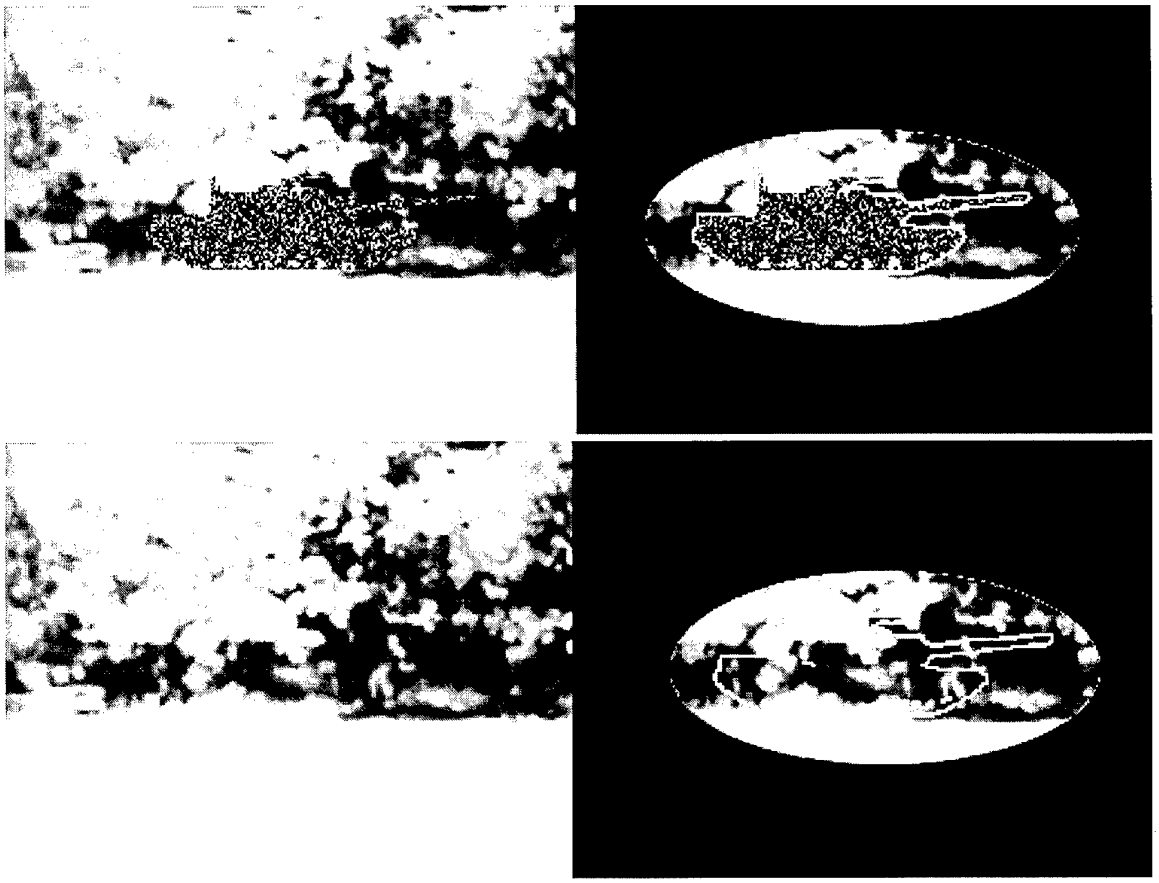


Figure 1 – Equal Signatures? Each of the pictures on the left shows a tank silhouette in a near-infrared scene (outlined to the right). The tanks have the same pixel intensity histograms and will give the same value for most signature metrics. Yet, psychophysically, these pictures are not equivalent.

The fundamental assumption of the LCJ is that when a person is presented with a physical stimulus, it elicits a psychophysical response, and that for any given stimulus, the response may vary from time to time and from individual to individual. Figure 2 shows a conceptual scale on which four stimuli (S_1 to S_4) have been rated. For each stimulus, there is a distribution of responses, which has been assumed to be Gaussian. When the psychophysical values of two stimuli are sufficiently close together, their distributions will overlap as shown in the figure. Under such conditions, it will happen that, for example, S_1 will sometimes be judged greater than S_2 on the psychophysical scale, even though it is actually less. This is called an inversion. It is important to remember that inversions are not "errors" in the normal sense, but the result of random fluctuations in the relationship between physical stimuli and psychophysical responses. In the extreme, two stimuli may be so similar that people cannot distinguish one from the other. In such a case, we would expect that in a forced choice situation, people would be approximately equally likely to pick each of the stimuli and the probability of an inversion would be approximately 0.5.

The LCJ is applied to data from paired comparisons in which people are asked to choose the stimulus that has the greatest (or least) amount of some attribute. For example, tones can be presented in pairs and the subjects could be asked which is loudest (or softest), higher (or lower) in pitch, shorter (or longer) in duration. Pictures can be presented in pairs and the subject can be asked to choose the one that is most beautiful, most relaxing, most representative of a place they would like to be, and so on. Samples of handwriting can be presented in pairs and the subjects can choose the one that is the most readable.

There are many means of ranking stimuli. However, for any given pair of stimuli, the LCJ permits one to do much more than determine which stimulus has most of the attribute being judged. From the amount of overlap in the distributions (represented by the probability of an inversion) one can calculate the distance between the true psychophysical values, provided the stimuli are close enough together that inversions are not too rare. Thus, inversions are a necessary feature of LCJ data, without which numeric scales cannot be ascertained.

As indicated above, people could be given many different tasks for the same set of images. If people were asked to choose the picture that represented the place they would most like to be, we would expect to get substantially different results than if we asked them to pick the one that was the most depressing. Thus, instructions given to the subjects define a task to be performed and greatly affect the choices that are made. Similarly, if we ask our subjects to listen to two tones and choose the one that is higher in pitch, even rudimentary musical training could substantially change the results. Clearly, then, the training and instructions given the subjects can greatly affect the outcome of an LCJ assessment and must be carefully controlled.

4. USING THE LCJ: A DEMONSTRATION

4.1. Procedure

We shall now demonstrate the use of the LCJ by applying it to a practical problem. This demonstration uses a set of 9 images from the Search_2 database[2]. The file names of the images

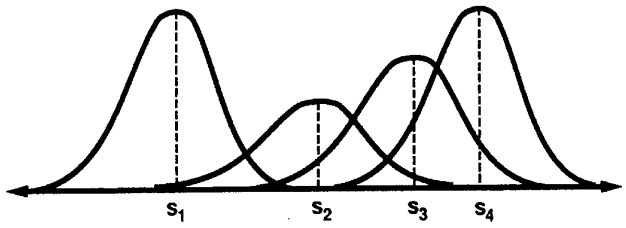


Figure 2 – A Conceptual Psychophysical Scale. This drawing shows 4 stimuli on a hypothetical psychophysical continuum. The horizontal axis indicates the amount of an attribute (e.g. beauty) that each stimulus possesses. The vertical axis indicates the probability that the stimulus will be judged to lie at that point on the continuum at any given time. The regions where the areas under the curves overlap indicate possible inversions.

and some of their statistics are shown in table I. These particular images were chosen because they represented a wide range of signature levels as indicated by mean search time, because they represented a small subset of the targets (all being T-72, M-3, or M-60), and because they represented a broad spectrum of probability of detection. As will be discussed later, it was necessary to keep the set of selected images small.

The images, which had been stored on a CD-ROM in photo-CD format, were read into Adobe PhotoShop® at resolution 5 (3072 x 2048 pixels) and printed 10.24 x 6.827 inches (26.01 x 17.34 centimeters) on 8.5 x 11.0 inch white bond paper using a Hewlett-Packard color LaserJet® 4500N printer.

The subjects (observers) were 13 engineers, scientists, and technicians who work with such images regularly in the context of search and target acquisition modeling and psychophysical evaluation. Prior to giving the images to a subject, the images were sorted into order by image number as indicated in table I. Each subject was told to re-sort the images into order from the one in which the target was easiest to find to the one in which the target was hardest to find. The subjects were not immediately told where the targets were in the images, but they were told that information was available when they wanted it. The results of their sorting are shown in table II.

As previously mentioned, LCJ analysis is performed on data from paired comparisons. Furthermore, it is necessary that every stimulus be compared to every other stimulus. Thus, for n stimuli, the total number of comparisons is

$$ {}_nC_2 = \frac{n(n-1)}{2}. \qquad (1) $$

Since this number grows much more quickly than n, it is necessary to keep the number of stimuli in any measurement block relatively small to avoid fatigue among the subjects and to keep the quality of their responses high. At the same time, since inversions are necessary, it is important that stimuli not be too far apart on the psychophysical continuum. While it is possible to obtain meaningful results with as few as 5 well-chosen stimuli, most practical applications limit the number of stimuli to somewhere between 10 and 25.

It was assumed that the subjects' judgements in a paired comparison evaluation would have been entirely consistent with their image collation order. Thus, it was assumed that any image in the sorted set would have been judged more difficult than any preceding image and less difficult than any later image in the set. This assumption was made because it is statistically most likely, even though inversions (inconsistencies) are common in practice. On this basis, each subject's ordering of the images was converted to a matrix in which a 1 in the i–th row and the j–th column meant that the i–th image was judged easier than the j–th image. Similarly, a 0

Table I. – Statistics for Selected Search_2 images

Image	Search Time		Nat. Log of Search Time		Visual Lobe		Correct Responses	Search Difficulty (LCJ)
	Arith. Mean	Geom. Mean	Arith. Mean	Geom. Mean	Detect	Identify		
Img0001	14.6	10.1	2.6810	2.3125	0.84	0.06	52	1.6480
Img0013	3.7	3.1	1.3083	1.1314	1.72	1.16	62	0.0000
Img0015	12.4	9.6	2.5177	2.2618	0.29	0.14	36	2.1964
Img0021	15.1	10.9	2.7147	2.3888	1.71	0.29	48	1.3143
Img0022	25.6	21.6	3.2426	3.0727	0.31	0.09	40	2.0914
Img0031	3.5	3.1	1.2528	1.1314	1.65	1.08	62	0.0000
Img0039	34.9	31.6	3.5525	3.4532	0.14	0.07	9	2.4224
Img0042	5.8	4.9	1.7579	1.5892	0.35	0.35	62	0.4920
Img0044	10.6	7.6	2.3609	2.0281	0.27	0.27	57	1.2000
R =	0.848	0.801	0.934	0.930	0.673	0.883	0.842	1.000
R^2 =	0.719	0.641	0.889	0.865	0.453	0.780	0.710	1.000

Table II. – Image Collation Order (Raw Data)

Person	Easiest								Hardest
DT	31	13	42	21	44	15	22	1	39
JeO	31	13	42	1	21	44	22	15	39
BB	31	21	44	42	13	22	5	1	39
DB	1	31	13	42	44	39	22	21	15
DW	31	42	13	21	44	1	22	39	15
JP	31	13	21	42	22	44	39	15	1
GO	31	44	21	13	42	1	15	39	22
JnO	31	42	13	21	44	1	9	22	15
KU	31	13	42	21	1	44	15	22	39
RD	31	13	42	44	15	22	21	1	39
MT	31	13	42	15	21	44	22	39	1
JK	31	13	1	44	42	39	15	22	21
MF	31	13	42	1	44	39	22	21	15

Table III. --TALLY MATRIX for subject DT.

1 ≡ first image was preferred. 0 ≡ second image was preferred.

	Second Image									
		1	13	15	21	22	31	39	42	44
	1	0	0	0	0	0	0	1	0	0
	13	1	0	1	1	1	0	1	1	1
	15	1	0	0	0	1	0	1	0	0
	21	1	0	1	0	1	0	1	0	1
	22	1	0	0	0	0	0	1	0	0
	31	1	1	1	1	1	0	1	1	1
	39	0	0	0	0	0	0	0	0	0
	42	1	0	1	1	1	0	1	0	1
	44	1	0	1	0	1	0	1	0	0

Table IV. --TALLY MATRIX from 9 images sorted by 13 people (Frequency of preferring first image).

	Second Image									
		1	13	15	21	22	31	39	42	44
	1	0	1	8	4	8	1	11	2	5
	13	12	0	13	11	13	0	13	10	11
	15	5	0	0	3	6	0	7	0	1
	21	9	2	10	0	9	0	10	3	8
	22	5	0	7	4	0	0	8	0	1
	31	12	13	13	13	13	0	13	13	13
	39	2	0	6	3	5	0	0	0	0
	42	11	3	13	10	13	0	13	0	10
	44	8	2	12	5	12	0	13	3	0

meant that the i–th image was judged more difficult than the j-th image. Table III illustrates this process, showing the matrix for the first subject listed in table II.

The matrices for all of the subjects were added, yielding the matrix in table IV. This matrix was the input to a computer program that applies the LCJ algorithms and produces scale values[3]. For the purposes of this paper, the program will be considered a "black box" with the details of the algorithms considered to be beyond the scope of the present discussion. The interested reader may wish to refer to Copeland and Trivedi[4], Torgerson[5] or Gescheider[6], or contact the author of this paper.

4.2. Results

4.2.1. LCJ Search Difficulty

The "search difficulty" values were calculated as described above and are included in the last column of table I. The last two lines of this table show the correlation (r and r^2) between the independent variable (LCJ "search difficulty") and the various dependent variables (metrics) that have been selected. One will observe that the search difficulty correlates very well with several of the metrics, particularly with the natural logarithm of the mean search time (either geometric or arithmetic mean). It seems appropriate to point out that scatter plots generally show very systematic relationships between the search difficulty and most of the selected metrics. However, some of the relationships are decidedly non-linear, causing systematic error when fit to straight lines. Thus, we find a substantially higher correlation between the search difficulty and the logarithm of the arithmetic mean search time (r = 0.934) than between search difficulty and the mean search time itself (r = 0.848). In the same way, the relationship between search difficulty and probability of detection is also non-linear (see figure 4). While table I does not have columns for the square and the cube of correct responses, the correlation coefficients are r = 0.923 for the square and r = 0.954 for the cube when compared to the search difficulty (LCJ). The graph in figure 3 shows the effect of search difficulty (as measured in this LCJ evaluation) on the logarithm of search time. This graph appears to be linear because the vertical scale is logarithmic. The graph in figure 4 shows the effect of search difficulty on the number of correct responses. The trend line shown is a quadratic function with r = 0.939.

4.2.2. Goodness of Fit

Testing the goodness of fit between the original data and the LCJ scale values (in this case, search difficulty) is a six-step process. One must first create a matrix D in which the diagonal elements are zero and for each off-diagonal element,

$$d_{i,j} = S_i - S_j \qquad (2)$$

where $d_{i,j}$ is the element in row i and column j, S_i is the scale value for stimulus i, and S_j is the scale value for stimulus j. Because the LCJ scale values produced above were chosen to use one unit normal standard deviation as the scale units, $d_{i,j}$ is the unit normal standard deviate for the separation of the stimulus mean response values. For example, since $S_1 = 1.6480$ (the scale value for Img0001) and $S_3 = 2.1964$ (the scale value for Img0015), then $d_{1,3} = -0.5484$ and $d_{3,1} = +0.5484$.

The second step is to produce a matrix Z in which each element $z_{i,j}$ is the predicted probability of choosing stimulus i over stimulus j. These probabilities are obtained either from statistical tables or by calculating

$$z_{i,j} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d_{i,j}} e^{-x^2/2}. \qquad (3)$$

Next, we calculate expected frequency of occurrence for choosing each stimulus i in preference to every other stimulus j. The elements of this matrix (E) are found by

$$e_{i,j} = Round\,(\, n_{i,j}\, z_{i,j}) \qquad (4)$$

where $n_{i,j}$ is the total number of times stimulus i is paired with stimulus j for all observers. Normally, this number is the same for all stimuli, in which case all $n_{i,j}$ can simply be replaced by n. For our example, the expected frequency of occurrence is given in table V.

The fourth step is to calculate

$$\chi^2 = \sum_{i=1}^{m} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}} \qquad (5)$$

where the values $o_{i,j}$ are the observed frequencies of occurrence from table IV. The upper limit of the summation is

$$m = \frac{k(k-1)}{2} \qquad (6)$$

where k is the number of stimuli in the experiment. However, the number of elements in the matrices O and E is k^2, and we are not using all of them, so it is necessary to define the selection process. In this case, we will select $o_{i,j}$ and $e_{i,j}$ only if $z_{i,j} \ni 0.5$. Furthermore, when $z_{i,j} = 0.5$, then $z_{j,i}$ is also 0.5 and $o_{i,j} - e_{i,j} = o_{j,i} - e_{j,i} = 0$. In these cases, we will use either of these differences, but not both. For our example, $\chi^2 = 16.3335$.

We shall next calculate ν, the degrees of freedom as

$$\nu = m - k \qquad (7)$$

where m comes from equation 6 and k is again the number of stimuli. In the example, $\nu = 27$.

Finally, the goodness of fit is determined by integrating the chi-squared distribution from 0 to χ^2 with ν degrees of freedom to obtain the probability of error. (The confidence is 1 minus the probability of error.) Normally one would not perform the integration, but use tables instead. However, the most common chi-squared tables in textbooks and most other sources only go up to 30 degrees of freedom. In our current, very limited case, $\nu = 27$. With 10 stimuli, the degrees of freedom increase to 35, and with 25 stimuli, it would be 275. It is clear that tables will normally not serve our needs.

There are at least two solutions to this dilemma. Available computer software can be used to calculate the probabilities. If you lack such software, the NCSS Probability Calculator[7] should serve your needs and is available free over the internet. Also if $\nu > 30$, the formula

$$d = \sqrt{2\chi^2} - \sqrt{2\nu - 1} \qquad (8)$$

may be used to calculate the normal standard deviate d associated with χ^2 and ν[8]. You may then refer to widely available tables for probabilities associated with the normal (Gaussian) probability density function. Such tables are found in statistics textbooks and standard mathematical tables. It may be sufficient to refer to table VI, which gives five key values of d, the probability of an error, and the corresponding confidence levels.

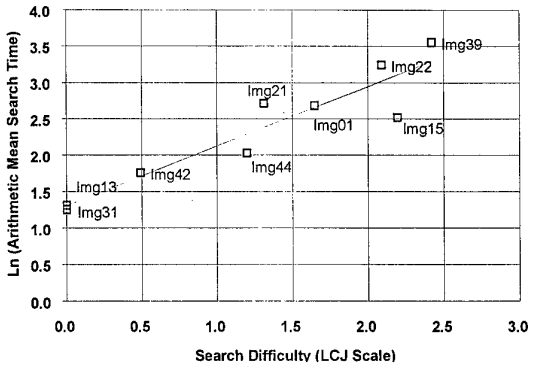


Figure 3 – Effect of search difficulty on search time. Nine images from the Search_2 database (r = 0.934)

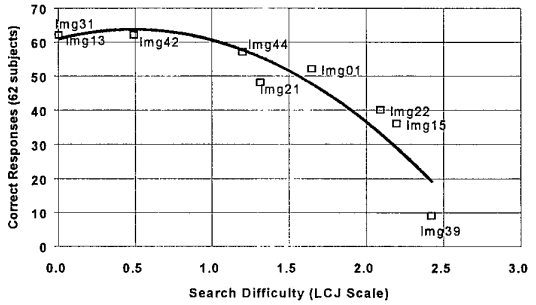


Figure 4 – Effect of search difficulty on correct responses.

Table V. – Expected Frequency of preferring first image (13 people).

First Image \ Second Image	1	13	15	21	22	31	39	42	44
1	0	1	9	5	9	1	10	2	4
13	12	0	13	12	13	7	13	9	12
15	4	0	0	2	6	0	8	1	2
21	8	1	11	0	10	1	11	3	6
22	4	0	7	3	0	0	8	1	2
31	12	7	13	12	13	0	13	9	12
39	3	0	5	2	5	0	0	0	1
42	11	4	12	10	12	4	13	0	10
44	9	1	11	7	11	1	12	3	0

Table VI – Probabilities associated with key values of the normal standard deviate.

D	Probability of error	Confidence
-1.282	0.10	0.90
-1.645	0.05	0.95
-2.326	0.01	0.99
-2.576	0.005	0.995
-3.090	0.001	0.999

Source: NCSS Probability Calculator

In the case of our example, equation 8 cannot be used because we have only 27 degrees of freedom. The NCSS Probability Calculator gives 0.054 for the probability of error and 0.946 for the confidence.

4.2.3. Repeatability

The group of observers in our demonstration sorted 5 of the images several days prior to the evaluation recorded in table II. The data was processed as described above and search difficulty values were calculated. When the scale values from the two sorting exercises were compared for these 5 images, the slope of the regression line was 0.997 and the correlation coefficient was r = 0.980. This indicates that the results were highly repeatable. However, since the process was not repeated with a different set of subjects, we cannot safely draw any conclusions about the performance of any other group of individuals or the population as a whole.

4.3. Discussion

The LCJ evaluation that was outlined above was relatively quick and easy compared to a properly run search experiment. At the same time, it correlates very well with search time and probability of detection. It would appear to be a highly effective tool for determining the relative strength of target signatures. At the same time, the LCJ has certain limitations.

The LCJ search difficulty scale that we obtained above is a psychophysical scale with no natural zero point and units that have no obvious relationship to useful quantities such as average time required to detect the target or probability of detection. Furthermore, the scale will change from one experiment to the next with no common reference. Thus, one might easily ask what advantage there is to such measurements. Is there any reason to use the LCJ in preference to other psychophysical measures or methods? I would like to suggest that there are numerous circumstances that might lead one to use the LCJ either in preference to other methods or in conjunction with them.

First, it is necessary to realize that the lack of a natural zero and a physically meaningful scale are really not significant problems. Detection time and probability of detection, while seemingly more meaningful are actually relative as well. The skill of the observers, the conditions under which the images are viewed, and many other variables in addition to the images themselves, will all affect the detection time and probability of detection. Observers who are more or less skilled, more or less effectively trained and motivated, or who are viewing images of varying quality and magnification will give varying results. Thus, in either case, two things are required: calibration standards and conversion formulas.

For example, in the case of the nine stimuli in the exercise above, the conversion from search difficulty to arithmetic mean search time in seconds can be expressed as

$$t \approx 5.21e^{0.82s} + 4.78. \qquad (9)$$

where t is time in seconds, s is the search difficulty, and e is the base of the natural logarithms. However, one must bear in mind that this formula applies only to the relationship between the search difficulty as measured by the data from the 13 Night Vision employees and search times for the 62 observers in the TNO test. It is likely that the 13 Night Vision employees could predict the search time on other images in the Search_2 set. It may also be that search times from the Search_2 data could be used to predict the search difficulty for other images. However, he who would extend this relationship to search difficulty values for other images sorted by other people or to search times in other search experiments would be making a potentially serious error.

Even so, all is not lost. Just as there was a day when two marks were scribed on a platinum-iridium bar to define a meter, other standards of measurement have been defined before and since. In the same way, useful perceptual standards can also be defined. However, rather than continue with search time and search difficulty, let us examine another phenomenon – visual clutter.

5. USING THE LCJ: A CLUTTER SCALE

The LCJ is primarily a tool for building measurement scales. Thus, we examine clutter as an example of an important quantity for which we have no accepted scale. Our purpose is to see how the LCJ could be used to build a standard reference scale. This relates to our primary topic of target signature evaluation in that target signatures must be evaluated in the context of a background and clutter is one of the most fundamental ways of characterizing backgrounds.

5.1. Definition of Visual Clutter

Clutter has been defined as "scene elements similar enough in size and contrast to the [target] that each one has to be considered in detail as a potential target"[9]. The concept of clutter is pervasive and generally describes distracting, annoying, and unwanted signals or returns when any of a wide variety of sensors is used. It is often discussed but seldom precisely defined. We shall use the phrase "visual clutter" in this paper to apply to any situation where there is a person using their eyes to examine a scene in which there is clutter, whether they are using "bare eyes" or an imaging sensor.

For many years, investigators have known that an observer's performance depends on many factors, including clutter. Schmieder[10] has probably been more influential than anyone else in the quest to subject visual clutter to quantification and analysis, but the proliferation of clutter metrics is testimony to the fact that none of the metrics are convincingly successful. However, the LCJ could be used as a tool in establishing a clutter scale that would be perceptually meaningful, extensible, and widely applicable.

5.2. Establishing a Unit of Visual Clutter

The first step in establishing a perceptual image clutter scale would be to select a set of images exhibiting a wide range of clutter levels. In order to maintain generality, they should represent numerous locales and clutter types. Since many feel that clutter must be understood in the context of the target, the set should include images with targets as well as images without targets. Initially, it would probably be satisfactory to have only military ground vehicles as targets.

From the initial set of images, a training package should be prepared so that observers can be taught what clutter is and so that they can become familiar with the size scales of the images in the set. This will help to make results repeatable, a necessary feature. A set of test stimuli would also need to be selected and should be distinct from the training set.

A pool of observers would also be required. The pool would need to be large enough that aberrant results from any one observer would have negligible effect on the results. Experience has indicated that at least 25 observers would be desirable. The observers would first be trained using the training set along with appropriate commentary. When they were fully trained, they would participate in a paired comparison evaluation of the test images. Their task would be

to choose the image in each pair that had the most (or least) visual clutter.

When all observers had completed the paired comparison evaluation of the test set, LCJ statistical analysis would be used to obtain the perceptual image clutter scale. At this point, the scale would be arbitrary. Probably the image that had the lowest clutter would be selected as the zero point.

From the test images, a subset would be selected as a reference set. Images that had the same, or nearly the same perceptual image clutter values would be culled. An attempt would be made to select a relatively small number of images that spanned the entire scale, and were evenly distributed between the extremes, but with no gaps. It would be best if about 1 unit normal standard deviation separated the individual images in the reference set. Probably 1 unit normal standard deviation would be selected as the scale unit.

It would be highly desirable to repeat the evaluation with a second pool of observers in order to establish whether or not the scale is indicative of a broader population. Actually, several replications would be ideal. If this could be done, the first replication should be with a group as similar to the first one as possible. Thereafter, greater liberties could be taken with the makeup of the observer pool in order to observe how robust the scale actually was.

5.3. Evaluating Clutter Levels

Having established a clutter scale for one set of images, one would naturally want to determine where other images were on the same scale. This could be done in any of at least 3 ways.

5.3.1. Quick Estimate

For a quick estimate of the clutter level in any image, anyone who was well versed in the perceptual image clutter scale could simply compare a new image to the reference images. Assuming there was nothing unusual about the image, they would be able to tell where it belonged on the scale, probably within about half of a unit. Tests of this method could be verified by one of the other methods to determine reliability.

5.3.2. LCJ Method

A second method of determining the clutter level in one or more images would be to mix new images with some or all of the reference set and perform a paired-comparison LCJ evaluation as described above. The results from the new evaluation would be used in conjunction with a linear transformation of scale values that would minimize the error for the reference images. This linear transformation could be determined by simply doing a linear regression between the standard values for the reference images and the values obtained for them in the new evaluation. The correlation coefficient obtained would be a measure of the reliability of the values assigned to the new images.

5.3.3. Jury Method

A third method would be to have a panel of "experts" who were all familiar with the perceptual image clutter scale assign clutter values to each of the new images. This would be more reliable and precise than the quick method above at the same time that it would be quicker and easier than the LCJ method. The major drawback to this method would be that there would be no ready means of evaluating the reliability of the values assigned to the new images.

5.4. Extending the Scale

If this methodology were employed, we might in time encounter clutter levels that were beyond the limits of the original set. There is nothing about the methodology in section 5.3 above that limits it to interpolation alone. In time, more reference images could be added to the set by the LCJ method outlined above. The only requirement in extrapolating beyond the original set is that no new set of images can be added if any continuous subset lies more than about one standard deviation beyond either end of the scale (depending on the number of observers in the pool). However, in such a case, selection of enough images with a variety of intermediate clutter levels should provide the necessary continuum.

5.5. Observer Pools and the Population

Near the end of section 5.2 above, we alluded to the fact that different populations might give different results. If this methodology were adopted for establishing a clutter scale, it would be wise to determine how stable the results were across these various populations. For example, it might be that trained military personnel would not give the same results as civilian clerical employees. On the other hand, since we are only asking individuals to make relative judgements ("Which image has the most clutter?") as opposed to quantitative judgements ("How much clutter does this image have?), we may find that the numbers obtained are quite stable over a broad spectrum of the human population. If the latter were true, it would be fortunate and knowing that it was true would permit various economies since trained military personnel are not always readily available at research facilities. At the same time, this cannot be assumed.

5.6. Analytical Methods

Naturally we would prefer to have analytical means of determining clutter levels rather than rely on psychophysical measures. However, we must remember that the human eye-brain system is most often the standard against which performance is rated. Having a reliable scale would be of great value in testing analytical methods because investigators would know what the "correct answers" are. Even if analytical methods were only able to tell which reference image a new image was most like, that would be a step in the right direction and eventually it could eliminate the need for paired comparisons and juries.

6. CONCLUSIONS

The Law of Comparative Judgement (LCJ) has great potential for helping us evaluate target signatures. There are two ways in which this potential might be realized. First, the LCJ can provide relatively quick, easy answers to questions that involve a complex set of variables such as we encounter when evaluating target signatures. It has been shown, for example, that the LCJ can give good estimates of mean search time using a methodology that is much quicker and easier that a traditional search experiment. When relative answers such as "Which is better?" and "How much better is it?" will suffice, or when there is a known relationship between LCJ scale values and important measures of effectiveness, the LCJ can be a highly effective tool. The LCJ can also be used to build scales for qualities that are difficult to quantify. This is perhaps where its greatest potential lies. To explain how this works, a scheme has been outlined for creating a perceptual image clutter scale. Such a scale could provide important benchmarks in an area of image understanding that has long

been in need of an anchor. Both of these applications could contribute greatly to the important area of target signature evaluation, search, and target acquisition.

7. REFERENCES

1. Green, D.M., and Swets, J.A., *Signal Detection Theory*, Peninsula Publishing, Los Altos, CA, 1988.
2. Toet, A., Bijl, P., Kooi, F.L., and Valeton, J.M., *A high resolution image data set for testing search and detection models*, (Report TM-98-A020), TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.
3. Copeland, A.C., "Xpet_pairs_LCJ" (Computer Program), Contract DAAK-70-93-C-0037, U.S. Army, CERDEC, Fort Belvoir, VA, 1997.
4. Copeland, A.C., Trivedi, M.M, and Ravichandran, G., *Developing a Quantitative Basis for Synthesis, Analysis, and Assessment of Complex Camouflage Patterns*, Contract DAAK-70-93-C-0037, U.S. Army, CERDEC, Fort Belvoir, VA, 1997.
5. Torgerson, W.S., *Theory and Methods of Scaling*, Krieger Publishing, Malabar, FL, 1985.
6. Gescheider, G.A., *Psychophysics: method, theory, and application*, Erlbaum Assoc. Publishers, Hillsdale, NJ, 1985.
7. *NCSS Probability Calculator*, Computer Program, NCSS Statistical Software, Kaysville, UT, 1995.
8. Weast, R.C, ed., *C.R.C. Standard Mathematical Tables, 13th Ed,* The Chemical Rubber Co., Cleveland, OH, 1964.
9. Lloyd, J.M., "Fundamentals of Electro-Optical Imaging Systems Analysis" in *The Infrared & Electro-Optical Systems Handbook, Volume 4: Electro-Optical Systems Design, Analysis, and Testing,* M.C. Dudzik, ed., SPIE, 1993.
10. Schmieder, D.E. and Weathersby, M.R., "Detection Performance in Clutter with Variable Resolution," *IEEE Transactions on Aerospace and Electronic Systems*, 19:4, 1983.

CAMEVA, A METHODOLOGY FOR ESTIMATION OF TARGET DETECTABILITY

Christian M. Birkemark
Senior scientist
Danish Defence Research Establishment
P.O. Box 2715
Ryvangs Alle 1
DK-2100 Copenhagen Ø
Phone: +45 3915 1746
Fax: +45 3929 1533
E-mail: cmb@ddre.dk

1. SUMMARY

This paper will present a methodology for computerised evaluation of camouflage effectiveness. The methodology is implemented in software at Danish Defence Research Establishment (DDRE) under the acronym CAMEVA. Basic input is a single image comprising a highly resolved static target as well as a proper amount of representative background. Separate target and background images can also be handled. Target and background regions are manually selected using the computer's standard pointing device (i.e. the mouse). From the input data, CAMEVA predicts the target detectability as a function of the target distance. The detectability estimate is based on statistical distributions of features extracted from the imagery, establishing a multidimensional feature space. In the feature space, the Bhattacharyaa distance measure is applied as an estimator of the separability between the target and the background. The intention is that the extracted features should resemble those applied during the human perception process. Typically, contrast and various measures of edge strength are applied. The Bhattacharyaa distance establishes a relative separability, while the absolute detection range is obtained by deriving a relation between the Bhattacharyaa distance and the estimated target resolution, at range. Thus by introducing parameters of the sensor, typically the human unaided eye, detectability as a function of the range is obtained. The methodology will not reflect individual observer performance but is aimed at providing an estimate of the optimal detection performance, given the selected set of features. During the choice of features and of sensor parameters, other perception mechanisms, than the human observer performance, can be modelled with this methodology. The paper will discuss theoretical and practical aspects of CAMEVA. Validation and application examples, including results on the NATO RTO/SCI-012 SEARCH_1 and SEARCH_2 datasets, will be presented together with other data.

Keywords: evaluating camouflage effectiveness, target detectability, detection range estimation, Bhattacharyaa distance, SEARCH_1 and SEARCH_2 data analysis

2. INTRODUCTION

CAMEVA is a methodology developed at the Danish Defence Research Establishment (DDRE) for computerised CAMouflage EVAluation and for estimation of target detectability. Input is a single digitised image comprising a highly resolved target as well as a proper amount of background. Based on that, CAMEVA predicts the target detectability as a function of range, in principle from relatively close range to infinite.

The methodology is based on measuring the dissimilarity between the statistical distributions of features on the target and on the background.

The need for objective and cheap methods for the evaluation of the effectiveness of camouflage measures was the original motivation for the development of CAMEVA. The methodology has however been applied in other tasks as well, including characterisation of sensors, evaluation of image processing algorithms and multiple sensor fusion algorithms.

2.1. History

In the late seventies and early eighties it was recognised at DDRE, that digital image processing was becoming available as a useful tool for solving many tasks of relevance to the military. Clearly it was important also to consider image-processing methods in relation to the evaluation of camouflage effectiveness.

The advantages were obvious. If such methods were available, the camouflage effectiveness could be assessed without the presence of costly equipment and significant numbers of human observers, as it is needed in field trials and in photo-simulation experiments. Evidently there were attractive economical aspects of this, but also the technical and scientific aspects were considered, such as the desire for speedy, reproducible results, that are easily documented.

For various reasons camouflage activities at DDRE were reduced to a minimum during a period of the late eighties and early nineties. The existing software was "mothballed" and further development within the regime of DDRE was cancelled during that period.

New tasks related to camouflage evaluation and to the estimation of detectability have however led to a resumed activity within this field. The old work was brushed up and that, together with new research in the field, led to the development and implementation of CAMEVA.

3. SYSTEM OVERVIEW

The aim with CAMEVA is to model the potentially achievable performance of a detection task. The aim is not to provide detailed models of the functionality of the very complex physiological and intellectual processes related to a human detection task. The aim is rather to establish limits to the performance that can be achieved, based alone on the information hidden in the scenario. Thus CAMEVA establishes an upper limit for the probability of detection.

A detection task is most often carried out by observers using the unaided eye, but this particular methodology is also applicable to observers using electro-optical equipment and as

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

a reference metric for optimality during the design of detection algorithms.

The high-level diagram of Figure 1 shows the main building blocks of CAMEVA.

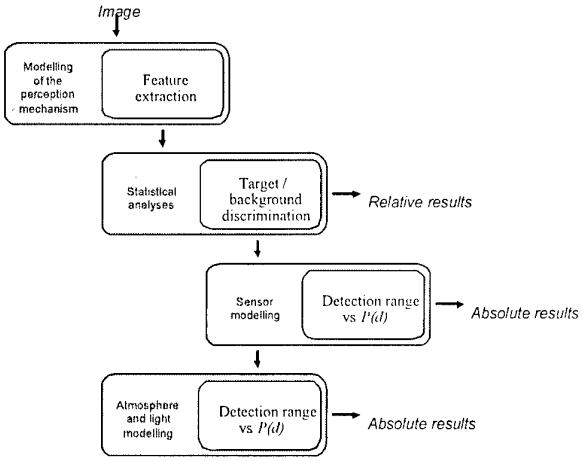


Figure 1: The main building blocks of CAMEVA.

The input is an image presumed to be of good quality. I.e. the image should have a high resolution compared with the degradations caused by the sensory system being modelled, and atmospheric transmission effects should be negligible. This assumption is needed due to the simulation strategy applied in CAMEVA. The results should be limited by the system parameters, and not by the quality of the input data.

The first processing block is the feature extraction. The target strength is based on the measurement of the strength of certain features on the target relative to the strength of these features on the background.

The choice of which image regions belong to the target and which regions belong to the background is based on an interactive selection procedure carried out by the system operator.

The choice of features depends on the characteristics of the presumed perception mechanism. If the unaided human eye is applied, features corresponding to those applied during the human perception process should be used. Or if some other perception mechanism is assumed, the features corresponding to that mechanism should be applied. A typical choice of features in the case of the unaided human eye is contrast, texture, shape, and edge-content sensitive features.

This is not a perception model in the more rigorous definition of this term, rather than it is a type of preprocessing attempting to extract the same attributes from the data, as those used during the processing performed by the actual perception mechanism. In the current context we will however denote this processing "the perception model" since this is the step in the processing where the characteristics of the psycho-visual processes are being modelled.

Based on the features, the analysis kernel estimates the statistical distribution of the features of the target and of the background. A measure of the difference between the distributions is established. This provides a relative measure of detectability, i.e. a measure that is independent of the target range. Finally the relative difference measure is weighted according to the degradation introduced by the limited resolution of the sensor and thereby also introduces the target distance as a system parameter.

Atmospheric influences on the results are currently not implemented. Consequently the underlying assumption is ideal atmospheric conditions. This is recognised as an important issue for the continuing improvement of the system. The idea is to propose an atmospheric weight factor to the measure of separability; similar to the way the sensor characteristics are modelled into the system. It has been demonstrated[2] that this, under certain conditions, can be done with a simplified transmission loss model.

The perception model typically used, is a simplified model of human vision, based on contrast and edge detection. The method applied in the analysis-kernel is based a statistical approach for measuring the separability between distributions, known as the Bhattacharyaa[3] distance.

The following sections will discuss various aspects of this methodology. The statistical decision theory involved is discussed initially, since this is also applied to illustrate the motivation for the choice of features. This discussion is followed up later, with considerations of atmospheric transmission and light conditions. Validation examples are shown at the end of the paper.

4. STATISTICAL DECISION

The statistical decision procedure is based on a feature vector $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. Each x represents one feature. We claim that detection is possible when the multidimensional distribution function of features on the target is sufficiently different from the corresponding background distribution.

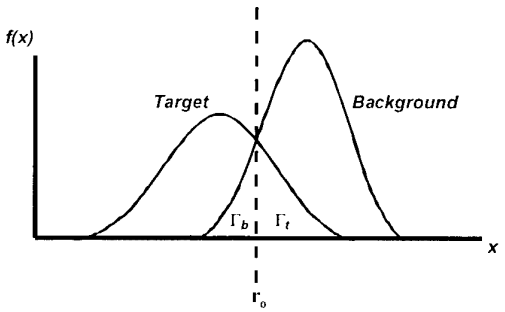


Figure 2: Distribution densities of target and background, with decision errors represented by Γ_t, Γ_b and the decision threshold r_o.

We further claim that estimates of the target and background distributions, based on high-resolution imagery, can be utilised to predict the corresponding distributions as a function of the distance. That is the same as postulating that basically the distributions are inherent properties of the target and the background. For the moment we shall neglect any external influence (i.e. atmosphere etc.) that violates this assumption.

4.1. Basic decision theory

The target and background distributions are exemplified for a single feature in Figure 2. In the detection case, assuming a one-sample observation, the fraction of the distribution overlap denoted Γ_t represents the error of missing a target with r_o as the detection threshold. Hence the probability of missing the target is $P(b)\Gamma_t$, with $P(b)$ as the a priori probability of a random sample being background. The errors represented by Γ_b (i.e. the false alarm rate) are irrelevant in this situation where we are evaluating a known target relative to the background. Correspondingly the probability of detection $P(d)$ is:

$$P(d) = 1 - P(b)\Gamma_t \qquad (1)$$

In general, with an n-dimensional distribution Γ_t is determined over the n dimensions of the observation vector.

By introducing the sampling and assuming k independent samples of the target, the detection probability becomes:

$$P(d) = 1 - P(b)\Gamma_t^k \qquad (2)$$

4.2. The Bhattacharyaa Distance

Given that parametric distributions are unknown, together with the general problems of analytically determining the distribution overlap $\Gamma=\Gamma_t+\Gamma_b$, we introduce the Bhattacharyaa[3] distance D, that is a measure of how different two distributions are:

$$D = -\ln\left\{\int_{-\infty}^{\infty}\sqrt{f_t(\mathrm{x})f_b(\mathrm{x})}d\mathrm{x}\right\} \qquad (3)$$

The integral is an approximation to Γ.

In the detection case where the target qua the distance occupies only a small fraction of the total field of view (FOV), a reasonable approximation to $P(b)$ is $P(b)\approx 1$. Furthermore it follows from $P(b)\approx 1$ that Γ_b is negligible compared to Γ_t and thus Γ_t is an approximation to the total distribution overlap.

Using $P(b)\approx 1$ and $\Gamma_t\approx\Gamma$, a minor rewrite of equation (2) provides:

$$P(d) \approx 1 - \exp\{k\ln\{\Gamma\}\} \qquad (4)$$

The distribution overlap is represented by the integral within the Bhattacharyaa distance, thus $D\approx -\ln\{\Gamma\}$. In the limiting case of identical distributions it is easily seen that $D=0$, and of totally different distributions that $D\rightarrow\infty$.

With the Bhattacharyaa distance we have:

$$P(d) \approx 1 - \exp\{-kD\} \qquad (5)$$

We see that in the case of identical distributions $P(d)=0$, which is reasonable due to $P(b)=1$. This means that there is virtually no chance of hitting the target by random choice. Similarly by totally different distributions $P(d)=1$, which is also reasonable since this allows detection of an almost infinitely small target.

4.3. Limited Resolution

We introduced k as the available number of independent samples of the distribution. In the visual detection task k will depend on the target size A, the target distance L and how well the eye is capable of independently sampling the FOV. The last parameter we shall denote "the effective minimum resolvable field of view" θ_k. The target size is normally trivial and the target distance is a variable parameter of the simulation (i.e. we are aiming at detection curves as a function of the distance). Thus with the assumption that θ_k is also known, the target resolution k is determined by the geometry of the scenario and we obtain the probability of detection as a function of the range:

$$P(d,L) = 1 - \exp\left\{-\frac{DA}{L^2\theta_k^2}\right\} \qquad (6)$$

Where k is:

$$k(L) = \frac{A}{L^2\theta_k^2} \qquad (7)$$

In the case of different horisontal and vertical resolution characteristics θ_k^2 is replaced by the product of the separated horisontal and vertical components $\theta_{k,h}$ and $\theta_{k,v}$.

4.4. Multidimensional Feature Spaces

Numerically computation of the Bhattacharyaa distance in the multidimensional feature space is quite complex and in fact not necessary. From the definition of the Bhattacharyaa distance we obtain with the assumption of independent features that:

$$D = \sum_{l=1}^{n} D_n \qquad (8)$$

We may therefore conveniently write D as the sum of the individual Bhattacharyaa distances for each feature.

This result is important since it implies that the complexity involved with the computation of estimates of multidimensional distributions can be avoided.

Whether this assumption is valid depends on the choice of features.

4.5. Feature extraction

Selection of features is a nontrivial task. This is illustrated in Figure 3. On the upper left "triangles" image, the two "targets" are easily detected visually. A statistically based metric, operating solely on the contrast, provides zero difference between both two targets and the background.

Clearly other features are also involved with the detection process in the example, although the contrast is still relevant. The three additional features in the Figure illustrates this:

It is evident that each of the features reflects different properties of the input data. We will consider the target and background regions of the added features:

The strength of the horizontal edges are applicable for detection of the upper right target (which we will denote "target A"), as these edges are relatively strong in the background, and virtually non-existent in the target. The lower left target (denoted "target B") is more difficult to detect, and the only useful cue is missing edge at the top of the target.

With respect to the vertical edges, target A is once again easily detected. This time due to the strong vertical edges.

Target B is difficult, although still detectable, due to the discontinuities of the edges. But basically the edge-strengths of both the target and the background are the same.

With respect to the isotrope edge features both targets are - roughly speaking - equally difficult to detect.

We see that the choice of features is vital to the result. In the case of an analysis (i.e. an algorithm) applying only contrast processing, this single feature might be sufficient. But in other situations more complex features should be applied.

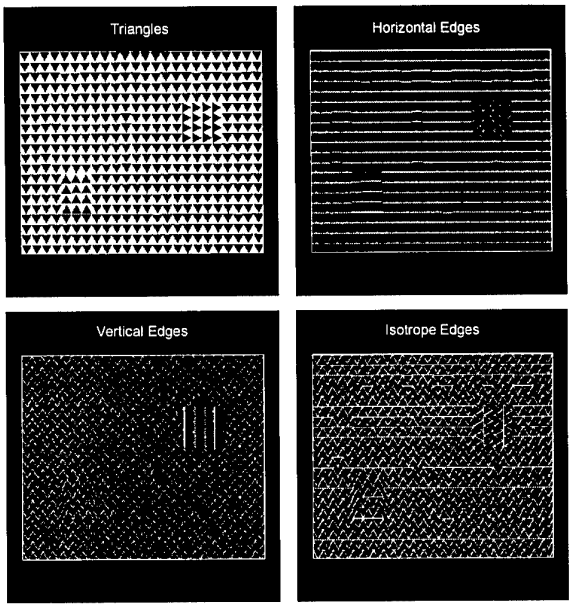


Figure 3: Four different features of a simple image with two target. Both the targets and the background are triangles. The unprocessed (the contrast feature) and three additional features based on edge strength.

By application of the relative analysis-part of CAMEVA, (i.e. computing the Bhattacharyaa distance), to the triangle input image with all four features, the computational results should reflect the qualitative assessment above. The results are shown in table 1.

The Bhattacharyaa distances correspond nicely to the results obtained from the visual assessment: D_C is about the same for both targets, which means that based on contrast both targets are equally easily detected. D_H and D_V each are relatively strong in target A, while D_H is weak in B. The isotrope component of the target is about the same for both targets, which is also reasonable since the have equal components of diagonal edges.

Typical features that have been applied in the application of CAMEVA to visual detection tasks are the four features used in this example.

4.6. Atmospheric and light conditions

While $k(L)$ provides a method of limiting the resolution due to distance, it does not model the effect of atmospheric transmission loss and light conditions (i.e. day or night). The current work will in principle assume infinite visibility and full daylight. A simple model for the atmospheric influence on the Bhattacharyaa distance have been considered[2], providing a Bhattacharyaa distance as a function of L. Likewise a simple model for the influence of the light is considered. Non of these modelling concepts have however been validated.

4.7. Training of algorithm

Training of CAMEVA is in principle a very simple procedure. There is only one parameter (θ_k) that is not easily determined, neither by the physical setup of the experiment, nor by the optics of the sensor.

The parameter θ_k depends on psychological factors such as motivation and on experience with visual target acquisition.

Table 1: Target/background Bhattacharyaa distances for the four different features of the triangle image.

Target	Feature			
	Contrast $[D_C]$	Horizontal edge $[D_H]$	Vertical edge $[D_V]$	Isotrope edge $[D_I]$
A	0.0336	0.1129	0.0894	0.0666
B	0.0363	0.0140	0.0501	0.0556

Based on the physiology of the eye[1], figures in the order of 0.2 mr are obtained. In the detection task however, where the observer is unable to focus on the target, this figure is typically too optimistic. Experimental data[1] prepared at DDRE have shown that θ_k=2.0 mr is a useful value for a group of non-military observers. In the case of highly trained observers, other values of θ_k may have to be used.

4.8. Practical aspects

To illustrate that the methodology applied in CAMEVA is quite widely applicable, we will discuss a few aspects related to the practical use of CAMEVA. This illustrates however also that this technique requires a skilled operator, with a general understanding of the problems related to the effectivity of camouflage measures.

- Data collected with detectability as the main purpose are typically long-range images. Often data taken at close range and with the target in the actual background have not been collected. In those cases experiments[6] with the application of CAMEVA have been made with separated target and background images. Clearly there are problems related to that procedure, but attempts have been made to align the internal target image-dynamics as well as the target-background image-dynamics in this type of data, thus allowing the target and background statistics to be taken from different images. Preferably however this procedure should be avoided, and should also be avoidable in experiments designed specifically as input to CAMEVA.
- CAMEVA is a human-in-the-loop process. With the input imagery, the operator must decide which image regions to apply as target and which to apply as the background. An example of this procedure is illustrated in Figure 4. Clearly that procedure is sensitive to the selection of in particular the background region, which is not well defined, and the operator must be careful, in order to produce meaningful results. Related to that is

Figure 4: A subsection of an image of the SEARCH_1[5] dataset, with overlay of target and background regions applied by CAMEVA.

also the fact that the results are basically a kind of averaging across the target region and the background region. I.e. if the target contains isolated bright spots, acting as cues to the observer, the averaging is problematic.

- In a classical camouflage-scenario the most efficient camouflage blends the target into the local background. CAMEVA is basically designed with that scenario in mind. In other scenarios, for example desert with scattered trees and bushes, the most efficient camouflage will sometimes be one that most closely matches the target to the scattering structure. Thus analysis made against the local background in that case are meaningless. In those situations the background region used as input to CAMEVA must be selected from the relevant part of the images, and not from the local background, like for example regions with trees and bushes.

Other scenarios can be thought of where there are problems related to this kind of methodology. In most cases however a skilled operator will be capable of producing useful results.

5. RESULTS

Results based on three different datasets are presented. Initially we discuss the JPRIS dataset that was specifically collected as a means for testing and calibrating CAMEVA. Secondly we discuss the application of CAMEVA to the SEARCH_1 and SEARCH_2 datasets.

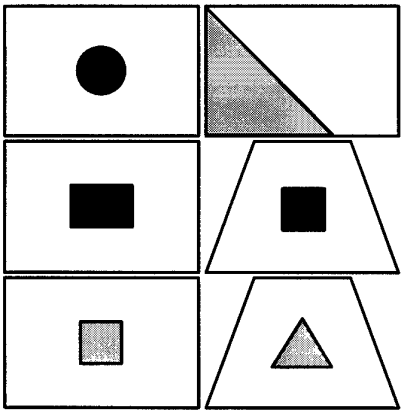


Figure 6: The types of panels applied as targets, not shown to scale. The panel areas vary from 0.20 m^2 to 0.90 m^2, while the signature elements vary from 0.025 m^2 to 0.26 m^2.

5.1. The JPRIS dataset

The jpris dataset is based on an observer field test[1] conducted by DDRE.

The experiment was designed as an observation trial with observations performed by the unaided eye.

Targets consisted of 12 artificial panels of 6 different types, either rectangular or trapezoidal as illustrated in Figure 6. Each panel was covered by a piece of green texture mat and a signature was attached to the surface, to allow identification when observed.

The background was a line of vegetation consisting of trees and bushes. All of the panels were presented against that same background.

The total length of the observation path was approximately 600 m with seven observation posts from long range (665 m) to close range (100 m) arranged along the path.

A total of 40 observers recorded the point of detection and the point of identification, starting at long range and approaching the target panels along the observation path. The observers were "semi-trained" in the sense that they were scientific and administrative civilian personnel working for the Danish military and some had, due to their scientific duties, some experience with observation tasks.

To support the computational analysis slides were taken from the observation posts. Each target panel was analysed with CAMEVA and detection curves obtained. Similarly detection curves were extracted from the field trial results. Comparison of the results, as averages across the observer population, is illustrated in Figure 7.

It is seen that the theoretical results to a very high degree describes the results of the field-trial. From these results it seems indeed, that the human acquisition process, under certain conditions, can be described as a statistically process derived from first principles.

The parameter θ_k=2.0 mr is however determined from the same population as were used as observers during the experiment.

One important error source is not taken into account during this experiment, namely the presence of eye-catching cues in the background close to the target. It is believed that the unexpected result for target-panel two is due to an abnormal intensity distribution (very dark) in the immediate neighbourhood of the target. There is evidence within the data that indicates that the observers have detected the background, and not the target itself.

5.2. The SEARCH datasets

The SEARCH_1 and SEARCH_2 datasets comprise a total of 44 images of six military vehicles recorded at different target distances and with the targets presented against different backgrounds.

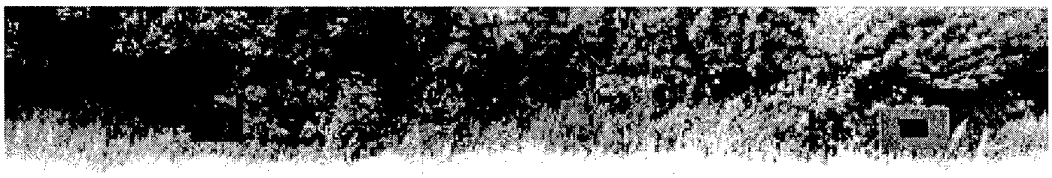


Figure 5: Example on the deployment of target panels in the JPRIS experiment.

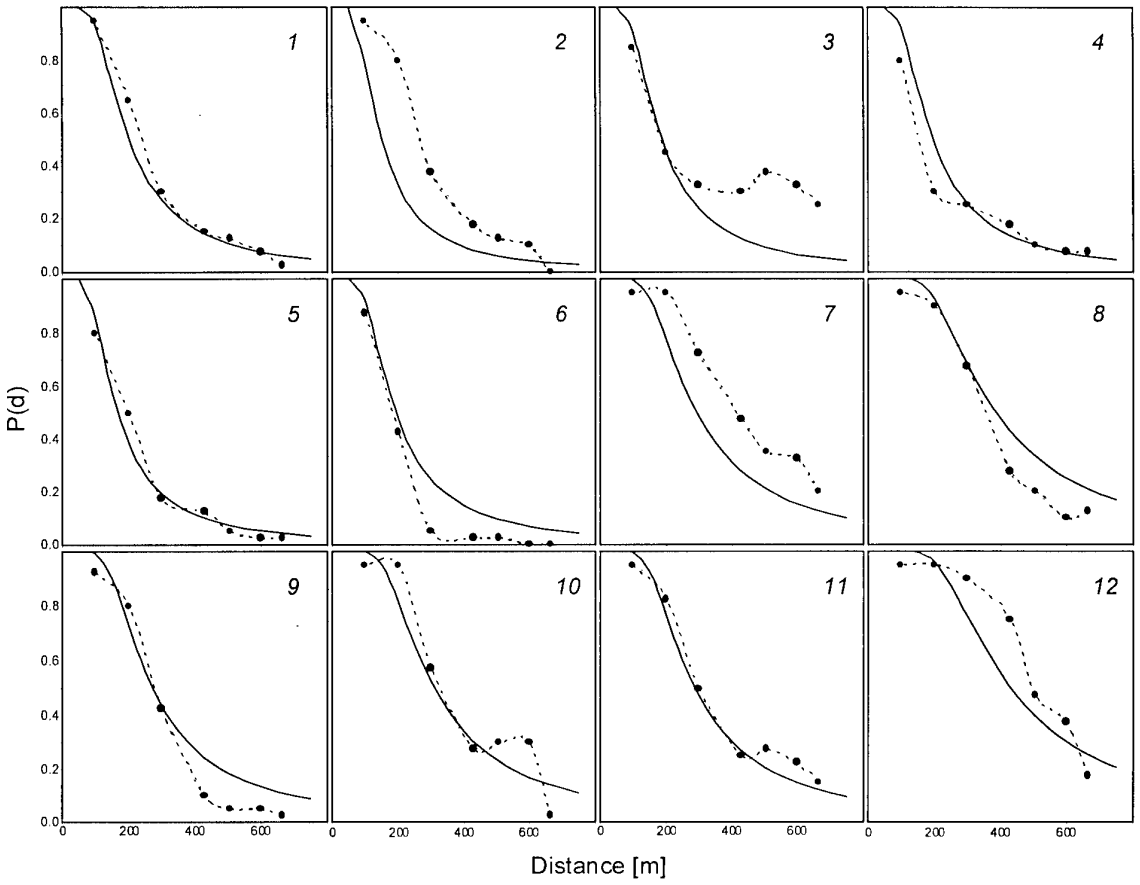


Figure 7: Detection curves obtained as a result of the DDRE JPRIS experiment compared with results of CAMEVA. Theoretical results are the solid curves, while the experimental results are the dotted curves

The only difference between SEARCH_1 and SEARCH_2 is the resolution of the digitised imagery, that is 1536 × 1024 pixels in the case of SEARCH_1 and 6144 × 4096 in the case of SEARCH_2.

Data are available in digitised form as full colour images as well as grey-level images. Each scenario is available as a pure background image, as well as an image with the target in the background. Furthermore close-up images of each target are provided at three different aspects. Ground truth is provided as binary target masks with each image. Target distances vary from about 800 m to 6 km.

The SEARCH data are collected and distributed by TNO-HFRI[1], and have kindly been provided for the application of the data to computational techniques estimating human observer performance in detection tasks.

TNO-HFRI has also conducted photo-simulation observer tests on the data, providing a baseline for the testing of computational methods for prediction of detectability and search time. A second photo-simulation experiment[7] on a subset of the SEARCH_1 data was prepared by DCTA[2]

CAMEVA was applied to the 44 images of the dataset and probability of detection was computed as a function of the target range. $P(d)$ at the actual target range was computed and compared with the photo-simulation results. CAMEVA was not trained on the SEARCH datasets prior to its application to them.

Cross comparisons of the results from CAMEVA, the TNO experiment and the DCTA experiment are sumarised in the bar-graphs of Figure 8.

As described CAMEVA normally provides detection curved as a function of the distance. In this case and to aid comparison of results $P(d)$ is computed at the actual target distance only.

6. CONCLUSIONS AND FURTHER WORK

Through the development and the study of CAMEVA, we have established a correlation between results obtained by human observers in the detection task and the results provided by CAMEVA. It is also evident however that several aspects of CAMEVA need further research to provide a more robust system. The most important areas are summarised below:

CAMEVA depends strongly on the skills of the operator during the selection of target and background regions. Rather than consider an automated procedure for that, which we believe is practically impossible, it is considered to produce a kind of catalogue that will set up typical scenarios together with proposed operator methodologies to cope with these.

[1] TNO Human Factors Research Institute, Soesterberg, The Netherlands

[2] Defence Clothing and Textiles Agency, Colchester, The United Kingdom

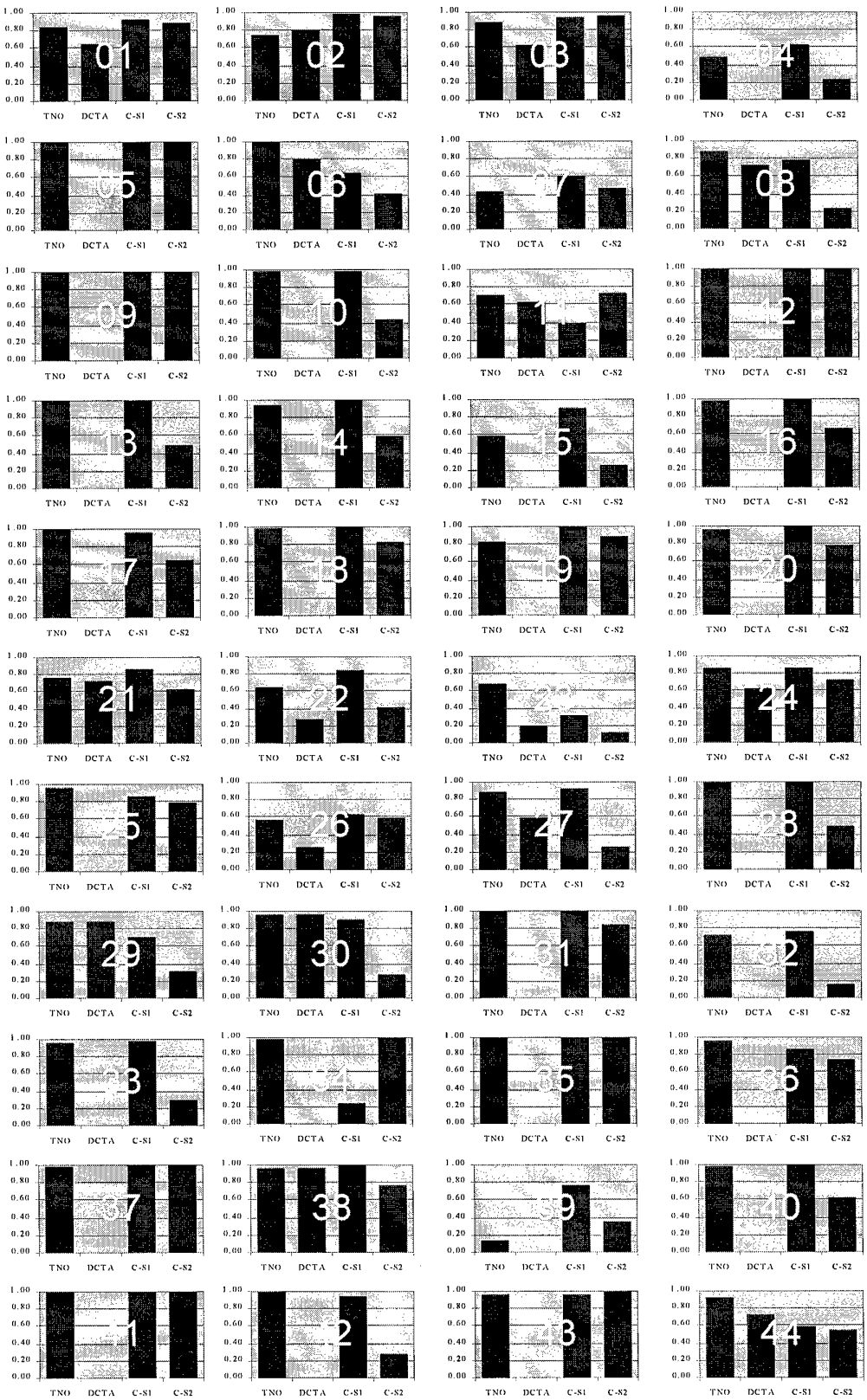


Figure 8: *P*(*d*) estimated by CAMEVA applied to SEARCH_1 (C-S1) and SEARCH_2 (C-S2), compared with observer experiment results.

It is considered important to implement a proper procedure for modeling of atmospheric transmission loss and of light conditions. Fundamentals for these sub-models have been investigated, but still need validation.

The current choice of features is not necessarily optimal. Certain aspects of detection are currently not modeled. A typical example is the cueing provided by long straight lines. Further features need to be investigated and in some cases algorithms for their implementation must be developed.

7. REFERENCES

1. Dannenberg, E. and G. Hvedstrup Jensen, *A Theoretic and Experimental Study of Human Visual Target Acquisition Based on Digital Image Analysis*, Danish Defence Research Establishment, DDRE 1982/19, 1982.
2. Hvedstrup Jensen, *Some Aspects of Statistical Separability and Detection Probability*, Danish Defence Research Establishment, DDRE 1981/30, 1981.
3. Fukunaga, Keinosuke, *Introduction to Statistical Pattern Recognition*, Academic Press, New York 1972.
4. Birkemark, C. M., *Computerised Evaluation of Camouflage Measures (CAMEVA),* Danish Defence Research Establishment, DDRE F-175, 1994.
5. Toet, A; Bijl, P; Kooi, F. L. and Valeton, J. M.: *Image data set for testing search and detection models,* TNO Human Factors Research Institute, Soesterberg, The Netherlands, TM-97-A036, 1997.
6. Birkemark, C. M., *Application of CAMEVA to the SEARCH_1 dataset.* Danish Defence Research Establishment, DDRE N-4, 1998.
7. Houlbrook, A., *Observer Tests on Digitised Imagery*, DCTA S&TD Research Memorandum 98/01, 1998.
8. Birkemark, C. M., *Theoretical and Practical Aspects of CAMEVA*, Danish Defence Research Establishment, DDRE N26, 1997.

EVALUATION OF TARGET ACQUISITION DIFFICULTY USING RECOGNITION DISTANCE TO MEASURE REQUIRED RETINAL AREA

Thomy Nilsson
University of Prince Edward Island
Charlottetown, Prince Edward Island, C1A 4P3, Canada
E-mail: nilsson@upei.ca

1. SUMMARY

The psychophysical method of limits was used to measure the distance at which observers could distinguish military vehicles photographed in natural landscapes. Obtained from the TNO-TM Search_2 dataset, these pictures either were rear-projected 35 mm slides or were presented on a computer monitor. Based on the rational that more difficult vehicle targets would require more visual pathways for recognition, difficulty of acquisition was defined in terms of the relative retinal area required for recognition. Relative retinal area was derived from the inverse square of the recognition distance of a particular vehicle relative to the distance of the vehicle that could be seen furthest away. Results are compared with data on the time required to find the vehicles in these pictures. These comparisons indicate that 1) the two methods are complementary with respect to distinguishing different degrees of acquisition difficulty; 2) recognition distance thresholds can be a suitable means of defining standards for the effectiveness of vital graphic information.

Keywords: vision, graphics, recognition, distance, retinal area, measurement, standards, target acquisition, TNO-TM Search_2

2. INTRODUCTION

2.1. Background

Graphic designers often contend that their work is too complex to be adequately represented by quantitative measurements of effectiveness. Yet the effectiveness of camouflage on military vehicles or the legibility of warnings on medications are examples where the effectiveness of graphic design has life and death consequences.

The power of the printed word led to a long history of quantitative research based on measurement of reading speed which has resulted in standards for the legibility of black letters on white backgrounds.[1,2] In an early attempt to extend such work to include color, Paterson &Tinker (1931) measured reading speed for words printed in various colors on various colored backgrounds.[3] They found that black/white were the most effective combination. A year later using the same colors, Preston, Schwankl & Tinker measured the effectiveness of colored print in terms of reading distance.[4] Blue/white, black/yellow, and green/white letter/background combinations were found to be the most effective. However, their results received little attention because subsequent replications using reading speed and recognition time continued to find black/white to be best while recognition time for other color combinations varied proportionally to lightness contrast.[2,5,6,7] This was seen as being consistent with vision theory that the middle and long wavelength based lightness-contrast mechanisms were the primary pathways for image detail.

Since color seemed to have no quantitative effect on legibility, its role in graphic design was presumed to be wholly a matter of aesthetic judgement. The results of Preston, et al. were conveniently forgotten. More recent contrary data on subjective legibility were ignored.[7] Also overlooked were data indicating different time constants of visual color mechanisms.[8,9] Convenience was probably another reason why most studies after the 1930's evaluated legibility using time measurements rather than distance.

Yet tachistoscopic presentations are not representative of reading tasks in the market place. Asked to measure the legibility of health warnings printed in color on tobacco packages, Nilsson & Percival reasoned that measuring legibility in terms of distance made more sense from a consumer's perspective.[10] However, greater distance did not adequately reflect subjectively greater ease of reading. Legibility was better described in terms of the required retinal image area based on the inverse square-root of distance. Subsequent research using distance to measure the effectiveness of foreground/background combinations of the six primary colors in messages, symbols, and outline drawings indicated that chromaticity contributed substantially to effectiveness.[11] This effect was not revealed when effectiveness was measured in terms of reading speed because chromaticity pathways are considerably slower than lightness-contrast pathways.

The effectiveness of camouflage depends on both color and pattern perception. Therefore, search time may not adequately reflect the contribution of color in recognizing such targets. The availability of the TNO-TM Search_2 dataset of high resolution images together with data on search time proved an opportunity to compare distance with time based measurements of visual effectiveness.[12] To help develop quantitative standards for more effective graphic design, this study evaluates data obtained by both methods.

2.2. Retinal Difficulty

The relationship between the distance at which a target can be recognized and the target's visual effectiveness is not as simple as might be supposed. When attention is directed at a target in a scene, the target's image falls on the foveal portion of the retina. Since the fovea has about one afferent neuron for every photoreceptor, the area of the target's image is proportional to the number of visual pathways available to convey information about the target. If all targets at

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

recognition threshold produce the same critical amount of information in the afferent pathways, a target's visual effectiveness or difficulty can be measured in terms of the retinal area needed for recognition. The retinal area required for recognition in turn depends on the amount of information per unit area of a target's image. We'll define a visually effective target as one that provides enough information for recognition in a small retinal area. Conversely, a difficult target is one that provides enough information only when its retinal area is large. At recognition threshold, actual retinal areas need not be calculated to compare the targets in terms of how much information they provide. The ratio of their effectiveness or difficulty depends only on the ratio of their threshold retinal areas.

The area of a target's retinal image is proportional to the target's size and is inversely proportional to the square of its distance. Hold target size constant for the moment. A measurement of the maximal distance at which a target can be recognized is inversely proportional to the retinal area needed for recognition. A long threshold distance means a small retinal area and therefore represents a visually effective target. Conversely a short threshold distance must represent a visually difficult target. The ratio of their effectiveness or difficulty can therefore be determined by the ratio of their threshold distances squared. Since the present research concerned measuring the effectiveness of camouflage, a ratio that reflects difficulty of recognition was used. The target that was recognized furthest away was taken as the standard. Its small retinal area was set to a unit value "1". The retinal areas of all other targets at threshold were scaled as multiples of this unit value and the result for each called its *retinal difficulty*.

In practice, these calculations were easy. Due to the inverse relationship between distance and area, retinal difficulty was obtained by dividing the threshold distance squared of each target *into* the threshold distance squared of the target that was recognized furthest. As an example, assume that a certain difficult target, X, had a threshold distance of 2 meters and that the least difficult target, Y, had a threshold of 4 meters. How much larger is the retinal area of X compared to Y? The retinal area of X is proportional to $1 / 2^2$; the retinal area of Y is proportional to $1 / 4^2$. In finding the ratio of these proportions, their *proportional-to-actual-retinal-area aspects* cancel, and the result directly equals the ratio of their retinal areas. Representing retinal difficulty of target X as R_X, the proportional retinal area of targets X and Y as A_X and A_Y, and their threshold distances as D_X and D_Y, we have:

$$R_X = A_x / A_Y$$
$$= (1 / D_X)^2 / (1 / D_Y)^2$$
$$= D_Y^2 / D_X^2 = 4^2 / 2^2 = 4$$

The more difficult target requires 4 times the area at recognition threshold than the least difficult target. Accordingly X's retinal difficulty equals the value "4" compared to target Y.

What happens when targets differ in size? The answer involves the concepts discussed so far, but also requires some additional concepts including: *visage* - target size with respect to a plane perpendicular to the viewer, *retinal information density* - amount of information per unit area of the target's retinal image, and *usable information density* - a quantity which reflects the limit imposed by visual acuity. Taken into consideration, they explain such obvious matters as why a large image can be visually effective even though it has a low information density, or why it is harder to camouflage a tank than a jeep. It was considered premature to deal with these concepts here. In the present study, retinal difficulty of seeing the targets was measured only in terms of image distance. Target area was taken into account using graphic analysis to reveal its effect.

3. METHOD

3.1. Subjects

Subjects were recruited by posters on campus and consisted primarily of psychology majors in their 2nd and 3rd years of study. They were screened for normal visual acuity using a Snellen chart and screened for normal color vision using the Dvorine Test. The purpose of the experiment was explained. They were asked to respond when they could no longer recognize the target in the picture before them as a vehicle while the picture moved away and to respond when they first recognized the target as a vehicle while the picture moved towards them. They were given several practice trials to get acquainted with using the controls and making judgements. Three females and two males viewed slide projected images. Four males and two females viewed the images on a computer monitor.

3.2. Apparatus

The subject was seated at one end of an 8 meter test track in a long, completely black, dark room. A carriage riding on linear-bearings either carried a Kodak Ektagraphic 35 mm slide projector with a 2.5 inch lens projecting 155 cm onto an HP rear projection screen or carried a Dell/Sony D1025, 17 inch, color monitor with 1280 X 1024, 0.25 pitch display. A computer-controlled stepping-motor accelerated and decelerated the carriage at 5 cm/s^2 or maintained a steady velocity of 10 cm/s. Carriage position was continuously monitored by an independent optical-encoder and electronic register.

Operational safety was ensured by program interrupts, limits set in the dedicated motor controller, and an independent system of limit switches that operated a clutch and brake on the chain drive. The computer also recorded the measurements, signalled when the image should be changed, and waited for the subject's instructions. Control buttons enabled the subject to direct the computer to start a trial, read the distance, or repeat the present trial. Images were changed manually by a researcher who was present at all times.

3.3. Images

A Polaroid Sprint Scan 35+ made 35 mm slides from a CD-ROM disk containing Toet, Bijl, Kooi, and Valeton's TNO_TM Search_2 data set of high resolution (3072 X 2048 pixel) images of various military vehicles in mixed rural landscapes of green foliage and pale yellow grass.[12] The scanner was calibrated in terms of a Kodak color calibration slide included on the CD-ROM. Accuracy of color reproduction was tested with a Topcon BM-7 colorimeter. The Y, x, and y values for the eight saturated color patches from brown to blue correlated +.60, +.97, and +.77

respectively with the Kodak values. Generally smaller y values presumably represented short wavelength absorption by the screen. Gray scale reproduction correlated +.92 with the values on the slide.

In many of the slide-projected images, the vehicle was difficult to discern even when the image was moved close to the subject. In a few images, the vehicle could still be recognized near the far end of the track. Eliminating these left 27 slides that were tested. Non-uniformity of brightness (mean = 89 cd/m^2, sd = 38) across the central portion of rear-projected screen was a concern since vehicle position varied considerably. Viewing the CD-ROM images directly on a computer monitor produced a crisper appearance overall, but the images were smaller than the projected images. Therefore these images were enlarged four times with the vehicle approximately centered on the screen. Twenty-eight of the most suitable ones were selected for testing with the monitor. The Y value of the saturated color patches on the monitor correlated +.99 with the Kodak values, but the x and y values could not be measured for most colors. Gray scale correlation was +.92 .

3.4. Procedure

The method of limits was used to measure recognition distance thresholds for the vehicles in the images. A within-subjects, ABBA counterbalanced design determined the order of image presentation. Each image was initially presented close to the subject. The location of the target vehicle was pointed out or verified with the subject. When ready, the subject signalled the computer to back away the image. When the subject could no longer recognize the target as a vehicle, he/she signalled the computer to record the distance. The carriage kept moving back a fixed plus random distance and was then brought to a halt. The procedure was then reversed with the carriage moving forward until the subject could recognize that the target was a vehicle.

Ten such measurements were taken in succession and ten thresholds calculated using running averages. The two thresholds that differed most from the mean were dropped and the mean and standard deviation of the remaining eight were recorded. All images were tested in single sessions that lasted between 90 to 120 minutes. Subjects rested between images, could take a longer break when they wanted, and were asked to rest a few minutes midway through a session. The two orders of presentation were tested on separate days. At the end of the last session, all subjects who viewed the monitor pictures were asked to look at each picture again at a distance about 0.2 meter and rate how difficult it was to see the vehicle using a ten-point scale.

4. RESULTS

4.1. Slide-Projected Images

Table 1 at the end of this paper provides the mean threshold distances and standard deviations for the subjects who viewed the slide projected images. *Up* and down refer to the order in which the images were tested. Results for each image were averaged across subjects. As explained in Section 2.2, each image's *retinal difficulty* was calculated on the basis of its mean threshold distance and the threshold distance of image #34, whose vehicle could be recognized farthest away.

Figure 1 shows retinal difficulty of the 27 slide projected images that were tested. The images are arranged in decreasing order of target size based on Toet et al.'s report.[12] Images were arranged in order of decreasing target size to reveal the effect of target area. For comparison, Toet et al's search time for these targets is also shown. Not surprisingly, both retinal difficulty and search tend to increase as target size decreases. Retinal difficulty correlated -0.69 with target size; search correlated -0.43. The correlation of the retinal and search results was +0.80. For the larger targets on the left, retinal difficulty increases faster with decreasing size than does search time. Compared to search time, retinal difficulty was notably larger for images number 20, 16, and 30. In each of these the vehicle is well outlined against its background and discernable features. Image 2 was noted for taking longer to find than might be expected from its retinal difficulty. Though the target in image 2 blends well with the background, the reader should bear in mind that its location was indicated to subjects doing the distance viewing.

For further insight into how the distance and time measurements differ, the effect of target size on difficulty was removed. The relative value of the reciprocal of target size was subtracted from relative values of retinal difficulty and search time. The results were then restored to retinal difficulty and search time values. Figure 2 shows retinal difficulty and search time results with the image area factor removed. Both functions have lost their generally upward trend as target size decreases. Yet various images such as numbers 16, 20, 30 , and 2 continue to differ substantially in recognition difficulty using either retinal or search measurements.

To help identify where the two sets differ, the images were arranged in order of increasing difference between the retinal and search measurements. These results are shown in Figure 3. This reveals that for most of the images there was marked correspondence of changes in both retinal difficulty and search time, even when the measurements differed somewhat. Comparison of those images that produced similar effects with those that differed revealed no systematic characteristics related to these trends.

Both the distance threshold and the search time measurements were tested for the significance of the differences between their means using Duncan's test. The results are shown in Figure 4. The means for distance threshold and for search time are arranged in decreasing order to represent increasing difficulty of recognition from left to right. The horizontal lines below or above the image numbers indicate the images that do not differ significantly. Generally, distance thresholds effectively distinguished targets that are easy to recognize while search time is poor at distinguishing these same targets. To a considerable extent the opposite seems to hold for images that are difficult to recognize. Of the 32 pairs of images whose search times did not differ significantly, only one pair, 10 and 18, had distance thresholds that did not differ significantly. Of the 21 pairs whose threshold distances did not differ significantly, four pairs, 10-20, 36-42, 38-32, and 29-26, did not differ in their search times.

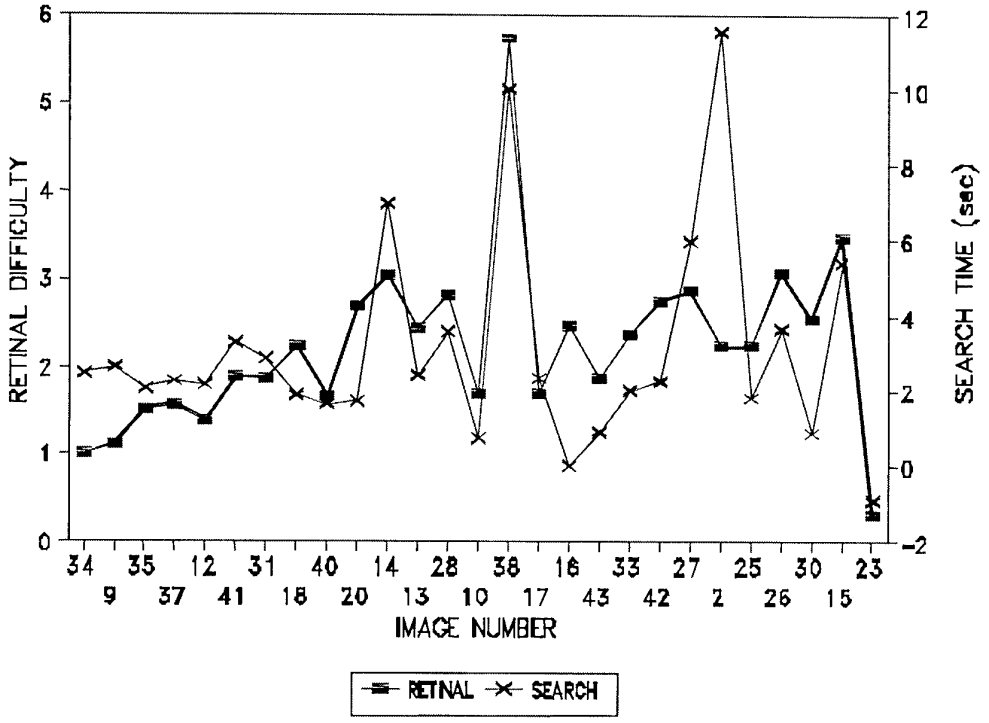


Figure 1. Retinal difficulty and search time of images that have been arranged in order of decreasing target size.

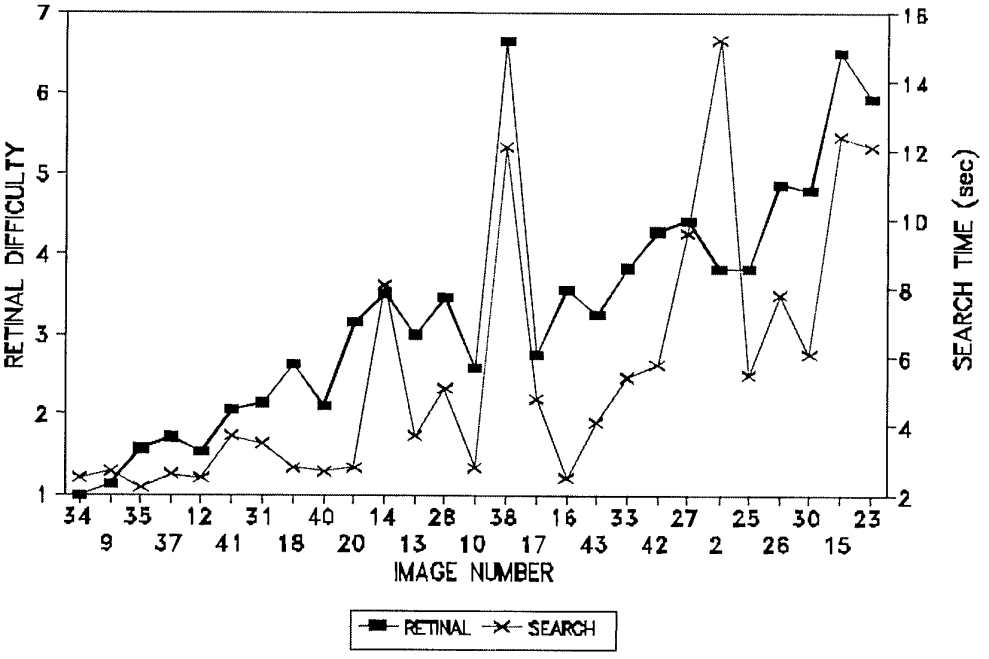


Figure 2. Retinal difficulty and search time with target size effect removed. The images arranged in order of decreasing target area.

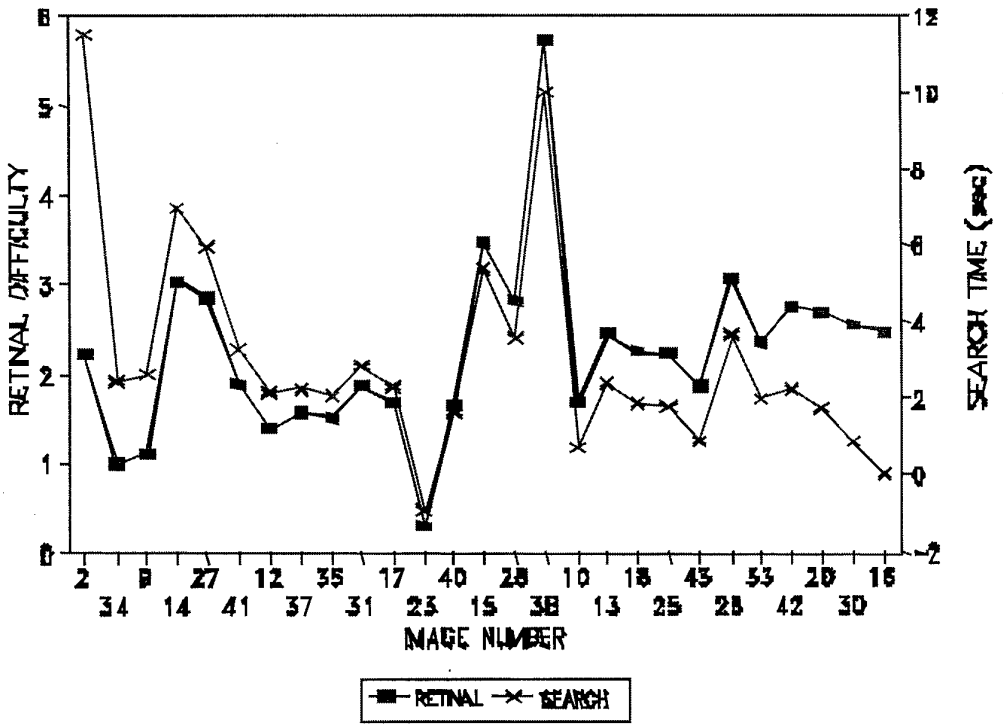


Figure 3. Images in Figure 2 rearranged to increasing difference between their retinal difficulty and search time

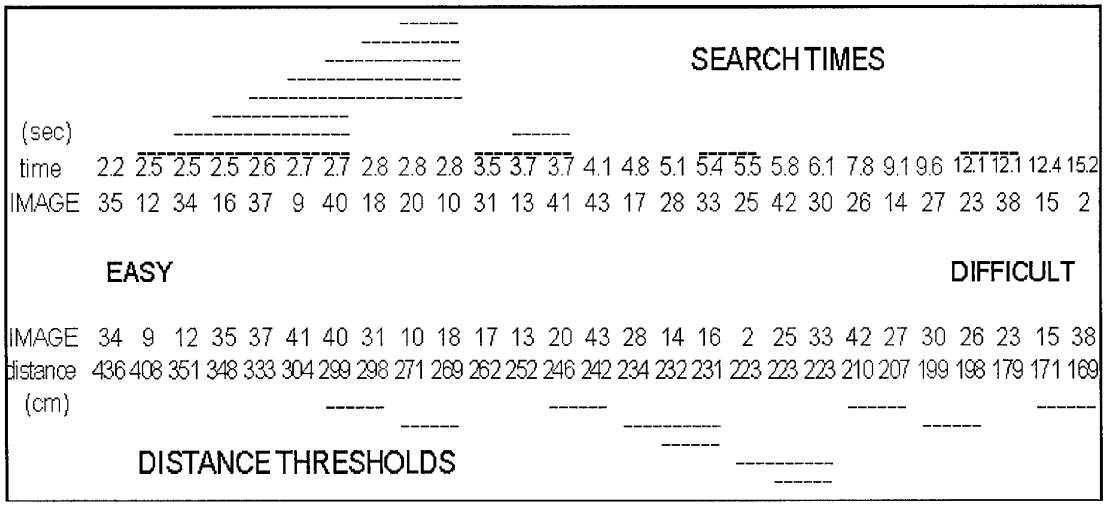


Figure 4. Results of Duncan's test on the significance of differences between mean threshold distances and between mean search times.

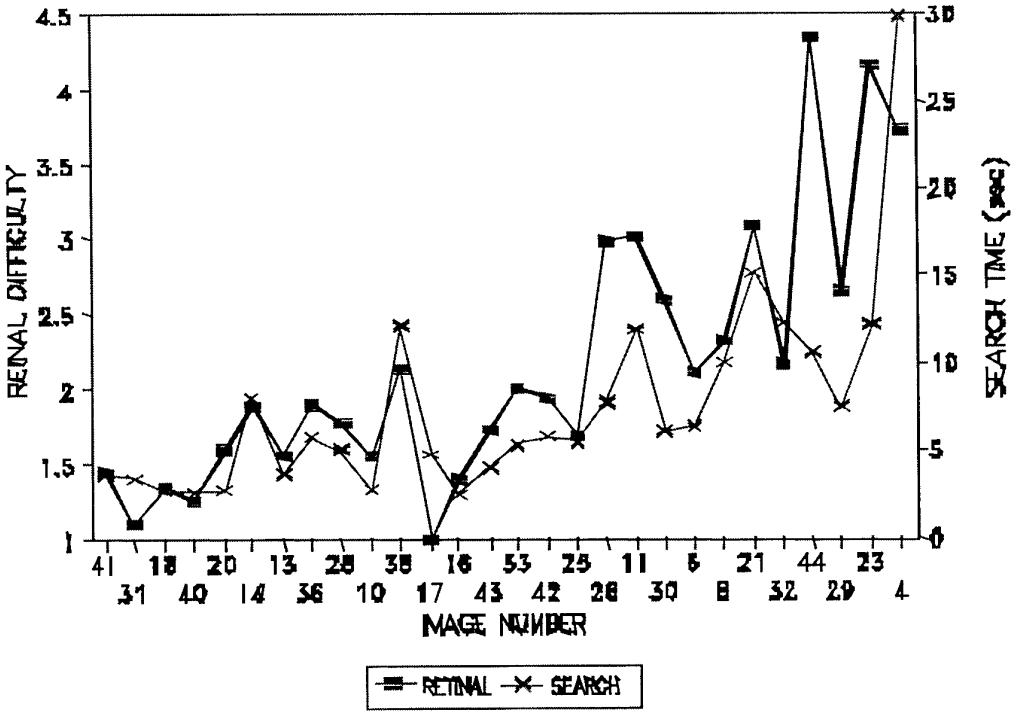


Figure 5. Retinal difficulty and mean search time for recognizing the targets in the various images presented on a monitor. Images are arranged in order of decreasing target size.

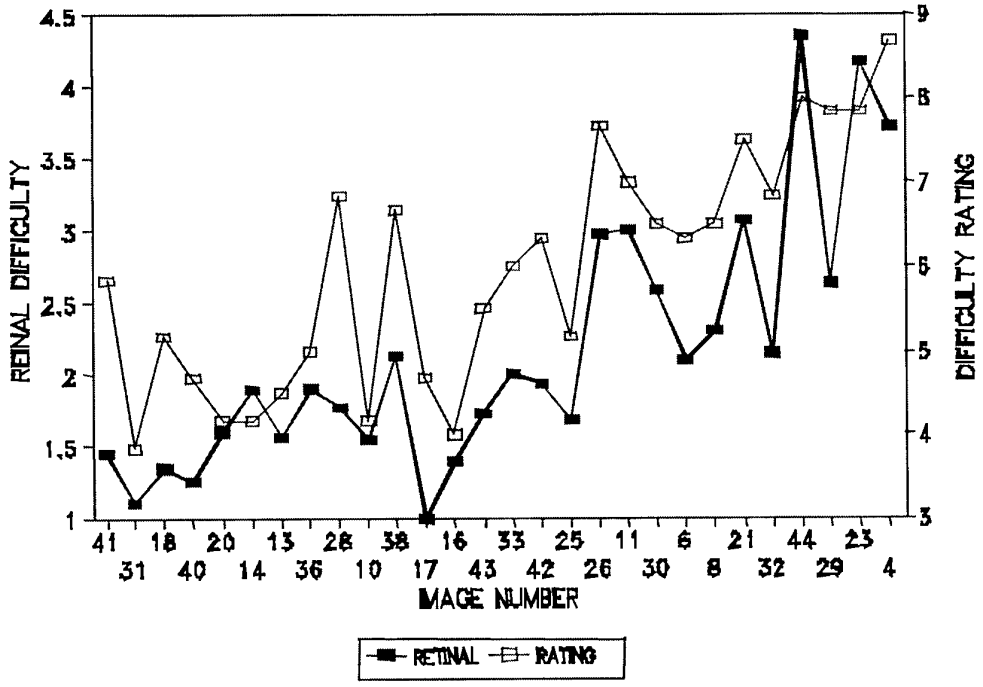


Figure 6. Retinal difficulty is compared with the mean rating of subjective difficulty of seeing the vehicle in each image. Images are arranged in order of decreasing target size.

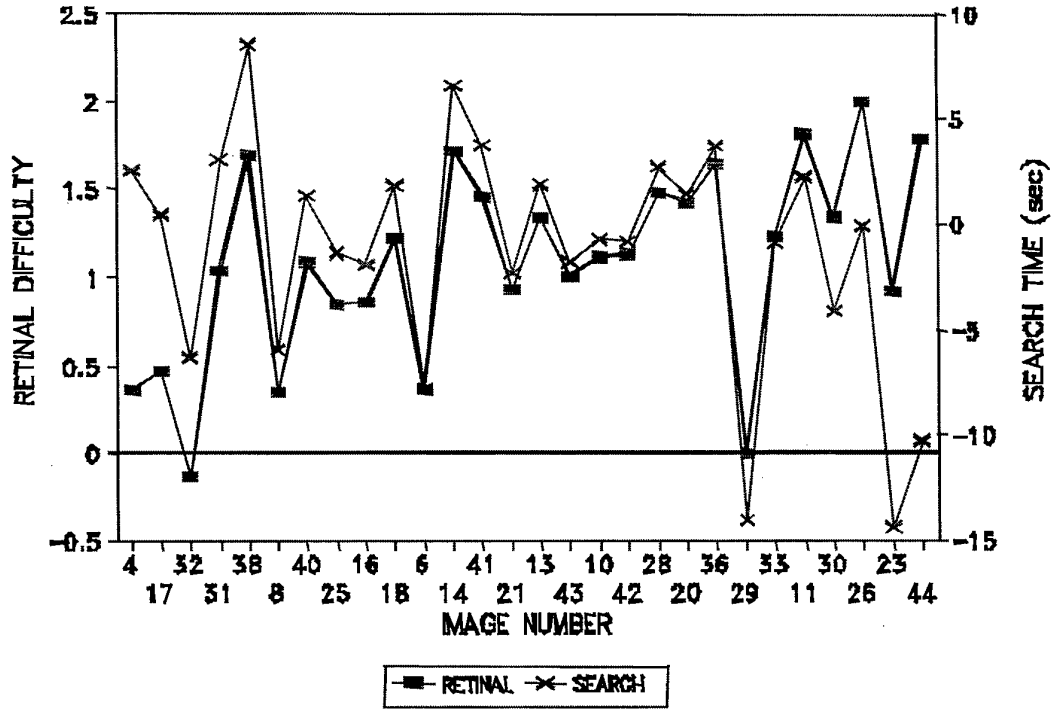


Figure 7. The effect of size has been removed from retinal difficulty and from search time. The results are plotted with the images in ascending order of difference between the two measurements.

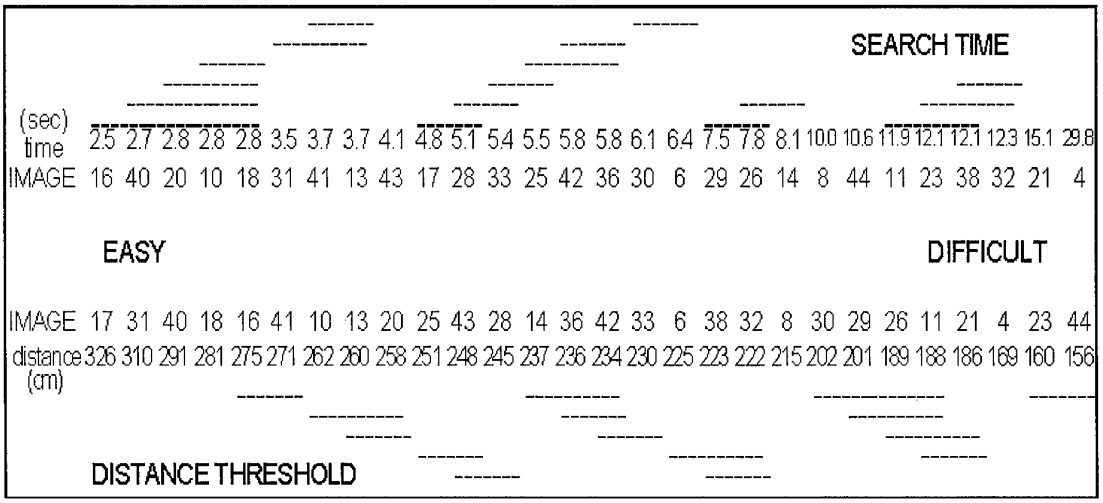


Figure 8. Results of Duncan's test on the significance of differences between mean threshold distances and between mean search times.

4.2. Monitor images

The threshold distance data for these images are provided in Table 2 at the end. As was done with the slide projected images, the distance thresholds for each of the 28 images tested were averaged across observers and converted to *retinal difficulty* values. The results are plotted in Figure 5 with the images arranged in descending order of target size. For comparison, the search time results measured by Toet et al. for these images are also shown. As target size decreased, both retinal difficulty and search time tended to increase. Retinal difficulty correlated -.61 with target size and search time -.49 with target size. As happened with the slide images, the two functions appear to track each other closely. Except for the smallest target, 4, subjects making distance threshold judgements had more trouble with the small targets (23, 29, 44) than those for whom search time was measured. The target in image 4 lacks both a distinctive shape, due to its head-on orientation and lack of contrast with its background. Targets 23, 29 and 44 have ample background contrast and have cues to their shape revealed by shadow and high-lights on the vehicles. Nineteen of these images were enlargements from the same images tested as projected slides. The correlation of the two sets of measurements was +.79.

For more information about the distance thresholds, retinal difficulty was compared with the ratings of subjective difficulty that were made by the subjects in the present study who viewed the images on a monitor. Figure 6 shows retinal difficulty and the mean ratings of subjective difficulty plotted as a function of decreasing target area. The two sets of measurements correlate +.85. The most notable differences again involve small targets in images number 23 and 44, which had relatively larger retinal difficulty values than subjective ratings. A similar difference was found above when the retinal difficulty of these targets was compared to their search times. On the other hand, the largest target in image 41 and a medium sized target in image 28 were rated notably more difficult than their retinal difficulty. While large, the target in image 41 is partly obscured by a tree and sits in the vicinity of other complex outlines produced by dark tree tops against pale grass. The vehicle in image 28 happens to line up with a light to dark transition across most of the scene. The subjective ratings correlated less well with search time, $r = +.72$ than did retinal difficulty, which is understandable since the latter involved the same subjects.

To see how differences between retinal difficulty, search time, and difficulty ratings were related to the size of the targets, the effect of target size was removed as was done with the data for projected images. Figure 7 shows the results for retinal difficulty and search time with the images arranged in order of difference between the two results. The biggest differences between the two types of measurement occurred for the images at each end of the graph depending on whether the search time (on the left) or retinal difficulty (on the right) had the larger value. As was noted when the size effect was not removed, the two types of measurement still differ most for images 4, 23, & 44. Removing the effect of size substantially reduced their correlation from +.72 to +.52. Removing the size effect did not change the correlation between retinal difficulty and the difficulty ratings.

Significance of the difference between means of distance thresholds of the images presented on a computer monitor was calculated using Duncan's test. The results are compared with those for the mean search times of these images in Figure 8. Of the 26 image pairs that did not differ significantly in search time, only 3 (42-36, 29-26, 38-32) did not differ in threshold distance. Similarly of the 21 image pairs that did not differ in threshold distance, 3 (10-20, 38-32, 29-26) also did not differ in threshold distance. Distance thresholds were somewhat better at distinguishing the easier images; search time was better for the more difficult images.

5. DISCUSSION

What do recognition distance thresholds reveal about the targets to be recognized? Target size had a substantial effect on both the distance and time based measurements of visual recognition. This can be seen in Figures 1 and 5 by the general increase in retinal difficulty and search time with decrease in target size. For the slide-projected images, retinal difficulty correlated -.69 with size, and search time correlated -.46; for the monitor images, the correlations were -.61 and -.49 respectively. The higher correlations of retinal difficulty with target size are understandable given the interaction between size, distance, and retinal area. While statistically significant (for 27 or 28 pairs, minimum significant $r = .38$) these correlations leave nearly 60% of the variance undetermined by the distance measurements. Some of that variance was due to measurement error. Despite the error arising from intra- and inter-observer variability, Duncan's tests revealed that a majority of the adjacent distance thresholds differed significantly. To what extent did these distinctions depend on target size which interacts with distance to determine the area of the retinal image? This is revealed in Figures 2 and 7, where the effect of size is removed from these measurements. Substantial variations in both retinal difficulty and search time as a function of the various images, indicate that other characteristics of the images influenced both measurement methods.

For additional insight into what other characteristics influenced these judgements, the effects of contrast were examined. The TNO-TM Search_2 data set provided measurements of the target vehicles' luminance and the surround luminance. From this a rough estimate of contrast was derived primarily based on the ratio between the dark area of the vehicle and the usually lighter grass. Retinal difficulty and search time correlated similarly with contrast (mean $r = -.35$ and -.37, respectively). Since these are not statistically significant correlations, this analysis indicates that it was not possible to estimate the contribution of contrast to recognition from these rough estimates. Research on colored symbols and backgrounds has found that chromaticity may contribute more to distance thresholds than does lightness contrast.[13] Chromaticity data were not available here, but their availability should be considered in future research.

Finally there is the effect of target shape. The vehicles generally appeared to have the most distinct shape when they stood sideways to the viewer. A rough estimate of the effect of shape was derived using the absolute value of a sine transformation of the vehicle's aspect angle in each image. For the slide-projected images, retinal difficulty and search time correlated -.39 and -.44 with this estimate of the shape effect; correlations for the monitor images were not significant ($r = -.25$ and -.28 respectively). The lower correlations of both retinal difficulty and search time for the

images viewed on the monitor were attributed to the different images in the two sets, since viewing conditions were constant for the search time data. The barely significant correlations with shape of data from the slide-projected images indicates that some of the differences in recognition were due to shape, but that aspect angle is not adequate for describing the effect of shape on these measurements.

How did the retinal difficulty measurements compare with the search time measurements? Evidence that the two methods were differently affected by target size is revealed in the results of the Duncan's tests in Figures 4 and 8. For both sets of images, threshold distances tended to distinguish larger targets better than smaller ones, while search times tended to distinguish smaller targets better than larger ones. Nevertheless, Figures 1 and 4 and show that both measurements of recognition difficulty generally increased with decreasing target size. This was also indicated by their significant, negative correlations with size as discussed above. However, many images were exceptions to this trend - some more so for retinal difficulty, others for search time. The two methods of measurement correlated +.80 for the projected images and +.72 for the monitor images. Since both measurements correlated substantially with image size, perhaps this explains their similarity. Removing the size effect reduced these correlations to +.61 and +.52 respectively. Note that the reduction was nearly the same for both sets of images. This similarity indicates that target size had a similar effect on both measurements. With the size effect removed, the results were arranged in order of the amount of difference between the two sets of measurements in Figures 3 and 7. In both figures, retinal difficulty and search time vary considerably as the images change. This indicates that differences other than the size of the images affected both methods of measurement. The close tracking of the two sets of results in both figures might lead one to think that two methods were similarly affected by these other characteristics. This was not the case. Reorganizing the results in terms of the differences does not change their correlations. Despite appearances, the pairs of functions in Figures 3 and 7 still have r values of only +.61 and +.52. Since Duncan's tests indicated a majority of the measurements were significantly different, the remaining variance was not all due to error. Evidently retinal difficulty and search time differed in how they were affected by other characteristics of the images.

Which method is better for distinguishing the difficulty of recognizing the vehicles? Duncan's tests indicate that threshold distances distinguished more images than did search time. For another way to address this question, subjects were asked to rate the overall difficulty of seeing the vehicle in each of the monitor-presented images. Figure 6 shows that the ratings increased more steadily with target size than did retinal difficulty. A comparison with Figure 5 shows this was also the case with respect to search time. Retinal difficulty correlated more strongly with the difficulty ratings ($r = +.85$) than did the search times ($r = +.72$). However, this is hardly a fair comparison since the search times were obtained from different subjects. Researchers studying the effectiveness of camouflage or other graphics should consider adding this simple measurement for comparison purposes.

Overall, the results obtained by Toet et al. and the present study indicate that the search time and retinal difficulty probably reflect different characteristics of the images. What does this imply for measuring the difficulty of recognizing camouflaged vehicles or, more generally, for measuring the effectiveness of graphics? Look again at the results from Duncan's tests in Figures 4 and 8. On average only 7% of the images that could not be distinguished by search time were also not distinguished by threshold distance, and 17% of the images not distinguished by distance were also not distinguished by search time. This suggests that the two measurements are indeed complementary. The use of both methods together would improve distinguishing camouflaged vehicles in terms of recognition difficulty. From a broader perspective, the favorable comparison with search time data on a standard set of images, provides further evidence that distance threshold measurements are an effective means of measuring the effectiveness of complex graphic displays.

6. ACKNOWLEDGEMENTS

This research was conducted in the UPEI - Health Canada Legibility Research Laboratory. Psychology student Donnay McNally collected much of the data. Thomas MacDonald of UPEI Audio-Visual Services transferred the TNO-TM Search_2 images onto 35 mm slides and enlarged the images for display on a computer monitor. Table 1. Mean distance thresholds (cm) and standard deviations for subjects viewing slide projected images.

7. REFERENCES

1. Sanders, S.S. and McCormick, E.J., *Human factors in engineering design*, McGraw-Hill, New York, 1993.
2. Tinker, M.A., *Legibility of print*, Iowa State University Press, Ames, Iowa, 1963.
3. Paterson, D.G. and Tinker, M.A., "Studies of typographical factors influencing speed of reading", *Journal of Applied Psychology*, 45, pp. 471-479, 1931.
4. Preston, K., Swankl, H.P. and Tinker, M.A., "The effect of variations in color of print and background on legibility", *Journal of General Psychology*, 6. pp. 459-461, 1932.
5. Knoblauch, K., Arditi, A. and Szlyk, J., "Effects of chromatics and luminance contrast on reading", *Journal of the Optical Society of America A*, 8, pp. 428-439, 1991.
6. Legge, G.E., and Rubin, G.S., "Psychophysics of reading IV. Wavelength effects in normal and low vision", *Journal of the Optical Society of America A, 3*, pp. 400-51, 1986.
7. Pastoor, S., "Legibility and subjective preference for color combinations in text", *Human Factors*, 32, pp. 157-71, 1990.

8. Courtney, S.M. and Buchsbaum, G., "Temporal differences between color pathways with in the retinal as possible origin of subjective colors", *Vision Research*, 31, 1541-1548. 1991.

9. Nilsson, T.H., 'The effects of pulse rate and pulse duration on hue of monochromatic stimuli', *Vision Research*, 12, pp. 1907-1921, 1972.

10. Nilsson, T.H. and Percival, T.Q., *Evaluating the legibility of tobacco health warnings with a method suitable for defining market guidelines*. (Report K302507), Health and Welfare Canada, Ottawa, 1989.

11. Nilsson, T.H. and Connolly, G.K., "Chromatic contribution to shape perception revealed in a non-temporal task using distance", in *Colour vision research: Selected proceedings of the International Conference, John Dalton's Colour Vision Legacy*, C. Dickinson, I. Murphy and D. Carden, Eds., pp. 197-206, Taylor and Francis, Bristol, Pennsylvania, 1997.

12. Toët, A., Bijl, P., Kooi, F.L. and Valeton, J.M., *A high-resolution data set for testing search and detection models*, (Report TM-98-A020), TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.

13. Clements-Smith, G., Nilsson, T.H., Connolly, G.K. and Ireland, W., *Development of an industrial service utilizing new technology to optimize visual information displays for both human and machine vision*, (Report 9F006-2-0021) Canadian Space Agency, Ottawa, 1993.

Table 1. Mean distance thresholds (cm) and standard deviations for subjects viewing slide projected images.

1st image	SM - up		JM -		GM - up		LH - up		PM - down	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
2	256	20.1	410	24.4	188	15.2	241	8.4	149	6.1
9	648	3.6	574	13.7	313	43.6	284	5.8	370	5.3
10	430	15.3	319	15.7	143	11.9	242	12.8	106	5.1
12	570	17.3	588	10.6	241	23.8	455	18.8	223	1.3
13	453	21.2	297	8.8	153	4.7	352	21.1	148	3.0
14	224	31.4	309	9.5	153	7.4	331	9.4	97	3.2
15	116	4.6	290	12.2	98	8.5	325	3.0	94	5.2
16	421	13.0	270	4.9	164	5.0	306	6.6	118	3.2
17	448	14.6	433	20.4	130	2.6	335	10.2	131	5.6
18	351	10.7	493	22.2	176	9.5	325	21.4	121	1.9
20	442	7.4	393	6.4	152	4.0	246	17.5	111	5.5
23	172	14.1	299	5.4	123	5.5	247	23.4	132	2.5
25	320	35.8	298	15.6	165	9.5	247	7.0	136	3.0
26	140	13.6	284	7.7	123	9.7	313	12.8	146	5.1
27	166	3.9	281	3.0	139	7.4	291	18.6	175	7.3
28	372	30.8	279	15.0	140	4.4	379	11.0	110	3.3
30	247	18.3	309	13.6	92	11.3	307	15.0	129	3.5
31	495	16.3	401	5.3	209	10.0	205	10.6	215	11.2
33	440	13.6	298	5.6	112	8.6	225	17.1	105	5.3
34	549	42.8	630	19.1	468	8.8	262	12.9	472	7.3
35	316	8.5	599	3.1	303	29.9	336	10.9	255	4.5
37	385	16.7	445	12.3	212	40.2	509	10.4	246	4.0
38	144	5.6	302	11.4	107	5.0	200	9.0	98	4.0
40	218	18.0	447	7.7	125	8.7	362	16.5	182	9.6
41	283	11.6	349	17.3	137	4.1	350	21.0	217	5.4
42	150	6.8	272	8.2	110	6.1	302	20.6	162	4.9
43	137	7.9	340	4.1	123	8.3	278	19.4	146	5.1
2nd	SM -		JM -up		GM - up		LH -		PM - up	
2	197	20.9	272	2.5	157	14.7	214	4.0	147	9.3
9	510	6.5	490	8.3	340	20.8	268	10.3	284	2.6
10	362	14.7	486	10.1	167	12.8	217	5.0	240	3.5
12	231	18.8	512	2.2	280	11.4	158	3.8	253	6.2
13	150	4.7	387	2.5	188	6.3	223	3.8	165	3.2
14	189	22.5	444	5.1	152	12.8	199	2.0	227	2.7
15	116	9.7	321	16.7	79	3.2	141	5.4	129	1.3
16	182	9.9	351	2.6	152	3.7	166	9.1	182	5.0
17	212	37.2	307	4.5	178	5.0	246	2.6	206	16.9
18	254	15.9	415	13.1	168	10.4	191	13.9	193	6.8
20	255	13.6	364	10.8	178	15.7	171	17.3	146	2.9
23	202	34.3	200	11.3	80	7.9	172	4.6	163	2.5
25	179	9.5	296	6.6	158	15.1	244	12.2	185	3.2
26	117	2.4	342	14.7	107	4.1	252	16.2	151	3.0
27	172	17.9	258	3.6	137	2.9	292	10.1	160	5.3
28	132	5.7	295	1.4	133	13.4	282	6.7	215	4.5
30	63	3.6	299	10.9	147	6.5	249	13.4	149	5.5
31	156	6.7	418	17.6	235	9.2	337	19.3	309	5.4
33	113	6.1	319	3.5	151	7.7	231	17.5	234	19.3
34	191	6.3	493	3.4	477	18.9	417	30.5	395	2.6
35	104	12.2	509	2.2	360	24.1	379	12.2	316	9.7
37	297	8.5	508	5.9	287	15.1	205	1.7	239	8.9
38	87	6.9	342	6.6	120	14.3	145	4.8	147	2.9
40	446	9.9	511	5.5	214	18.6	269	20.5	218	8.9
41	446	6.1	503	5.7	184	5.6	313	21.1	254	3.6
42	179	9.4	437	9.9	158	12.5	157	5.1	179	5.7
43	197	6.2	443	9.7	242	26.0	291	14.3	219	14.0

Table 2. Mean distance thresholds (cm) and standard deviations for subjects viewing monitor images.

1st	TC - down		RM - down		PM - up		GM - up		BF - up		SL - down	
image	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
4	83	10.0	107	4.2	109	3.2	172	9.1	149	4.3	220	8.6
6	88	2.4	169	17.1	190	2.7	251	12.9	259	9.5	217	22.5
8	93	7.1	200	6.1	191	9.7	229	11.0	205	8.1	167	6.6
10	98	3.5	158	12.7	234	9.6	332	23.7	232	12.2	277	17.2
11	73	2.8	94	7.8	155	1.8	307	30.5	187	10.3	176	9.8
13	112	2.0	181	19.0	208	8.4	298	18.1	203	8.2	248	13.0
14	69	3.2	110	23.9	367	11.1	250	15.0	222	13.5	213	9.7
16	94	10.1	302	33.5	168	12.5	347	24.3	216	7.9	308	16.9
17	132	6.8	397	20.7	199	6.7	376	34.1	267	22.2	289	19.4
18	96	5.0	199	15.3	233	8.4	339	14.4	238	7.4	277	7.6
20	114	18.9	129	13.0	250	3.1	365	6.7	245	15.8	281	11.2
21	63	4.5	142	5.1	198	4.8	263	22.8	144	6.6	166	0.9
23	67	3.8	124	7.2	145	11.1	235	18.5	182	3.1	91	11.3
25	125	8.8	139	11.4	344	10.0	299	26.5	205	10.2	173	7.7
26	62	1.7	127	16.9	207	2.8	226	25.2	160	4.3	187	11.8
28	82	5.2	165	26.0	275	8.8	251	16.1	170	4.3	250	7.3
29	71	7.9	122	6.9	205	12.3	187	9.0	165	5.2	225	9.0
30	78	2.3	163	25.8	190	6.2	255	25.2	157	1.8	230	13.6
31	196	9.0	217	19.6	358	14.1	428	48.9	233	13.2	289	17.3
32	78	9.5	143	7.6	287	1.3	222	14.1	200	5.7	257	8.3
33	100	6.1	188	19.1	198	7.7	256	4.8	229	15.3	297	7.3
36	133	11.7	152	16.5	263	8.8	232	6.4	255	8.3	234	15.4
38	120	6.3	115	13.6	175	5.0	229	16.0	256	10.0	230	6.0
40	105	6.1	215	6.7	376	14.3	262	32.5	271	17.9	357	24.6
41	153	18.1	167	20.6	350	9.4	264	13.0	219	11.4	338	4.6
42	121	8.5	145	16.4	249	6.1	233	12.7	194	6.7	307	13.4
43	167	12.3	134	13.2	265	6.2	270	33.0	187	4.6	323	18.2
44	81	4.9	89	3.6	190	3.7	150	11.8	81	4.9	137	8.7
2nd	TC - up		RM - up		PM - down		GM - down		BF - down		SL - up	
4	108	11.1	134	7.1	138	4.0	199	15.8	236	17.5	373	13.9
6	159	12.0	169	13.0	246	10.4	345	16.1	325	10.6	278	20.8
8	148	12.7	172	11.1	242	9.6	338	13.2	313	15.1	276	21.0
10	224	8.5	199	16.6	302	4.0	405	31.9	295	14.1	386	12.9
11	115	4.0	119	6.9	189	10.4	368	30.0	221	7.5	252	30.9
13	199	12.4	224	16.8	242	4.5	425	22.7	324	20.3	462	41.8
14	117	7.1	196	10.7	212	6.0	380	25.8	318	5.8	385	17.8
16	220	6.4	216	8.7	278	4.7	409	13.5	291	8.1	451	27.7
17	265	8.4	366	27.5	309	1.6	502	11.4	307	18.5	500	25.6
18	281	9.8	235	12.5	241	8.6	391	21.5	328	23.8	511	43.6
20	230	9.5	221	11.8	217	5.0	349	10.1	292	12.5	400	14.0
21	125	7.3	243	26.5	132	4.7	246	25.3	195	8.4	316	5.4
23	90	5.1	185	16.0	174	8.4	150	6.0	245	5.8	226	19.1
25	358	16.7	220	8.1	251	8.3	315	29.1	237	14.7	342	15.1
26	130	14.3	214	19.8	209	8.8	242	26.8	218	8.6	285	16.2
28	152	14.7	322	15.8	293	10.8	301	24.0	248	6.3	428	26.6
29	184	25.5	196	16.4	212	5.9	220	15.6	257	4.0	363	21.6
30	106	11.0	181	9.1	250	8.3	302	14.3	226	13.1	291	16.6
31	238	11.0	310	5.5	348	5.8	436	16.7	232	10.2	440	26.7
32	162	6.9	141	14.6	291	13.2	323	19.6	240	9.9	321	21.3
33	216	21.7	184	30.0	203	4.4	283	8.9	246	7.5	364	17.4
36	238	21.2	191	16.5	247	12.5	316	37.3	249	5.6	326	25.0
38	277	27.1	157	20.7	193	2.7	341	23.3	279	10.7	305	18.1
40	125	10.9	420	23.3	258	4.5	362	19.1	297	9.4	442	21.5
41	219	18.4	237	26.8	359	7.7	332	13.9	249	9.1	359	10.6
42	122	12.2	302	14.0	247	6.1	231	17.8	259	10.0	397	28.5

EVALUATING TNO HUMAN TARGET DETECTION EXPERIMENTAL RESULTS AGREEMENT WITH VARIOUS IMAGE METRICS.

G. Aviram, S. R. Rotman,
Ben-Gurion University of the Negev, Department of Electrical and Computer Engineering.
P.O. Box 653 , 84105, Beer-Sheva, Israel.Tel. (972)-7-6461518. Fax. (972)-7-6472949
E-mail: guyavi@ee.bgu.ac.il srotman@ee.bgu.ac.il

1.SUMMARY

An evaluation of the agreement between experimental results of human target detection performance, as obtained by TNO - Human Factors Research Institute, and various image metrics is addressed in this paper. Image metrics, such as local target from background distinctness metrics (*DOYLE* and *TARGET*), a global image complexity metric (*POE*) and a textural global / local co-occurrence matrix metric (*ICOM*), are presented and applied to the TNO image database. Good agreement, denoted by relatively high correlation levels, is found between the experimental results (search rates and probabilities of detection) and both *DOYLE* and *TARGET* local image metrics values. On the other hand, a relatively low correlation level is obtained between the experimental results and the *POE* global image metric values. Correlation values obtained using the global / local *ICOM* metric are between these extremes, as expected. These results emphasize the dominance of the target to background distinctness perceptual cue and the appropriateness of the local metrics to this kind of imagery. Furthermore, they can be used to formulate empirical classification rules that can be used to evaluate and predict human detection performance in similar cases.

Keywords: human target detection, clutter, contrast, texture, image metrics, psychophysical experiments.

2. INTRODUCTION

The Search and Target Acquisition Workshop is an excellent opportunity to deal with one of the most challenging topics in the field of human image perception. This topic is the evaluation of the relationship between image content and human target detection performance. The traditional attitude to deal with this challenge is to define image metrics to measure various kinds of perceptual cues, apply them to the natural scenes and correlate their outputs with target detection experimental results. The higher the correlation, the more appropriate is the corresponding metric. Following this attitude, much work has been done in recent years, mainly with infrared imagery, creating many image metrics of different kinds, all designed to emphasize one or more of the parameters that dominate target detectability by the human observer. Although this research greatly contributed to the understanding of human perception process, no unanimous decision upon the best metrics exists. The goal of this paper and its main contribution to the workshop is to evaluate the appropriateness of various image metrics, originally designed for infrared imagery, to imagery in the visible region of the spectrum. This task was fulfilled by applying our research methodology and analysis techniques[1,2], to the TNO image database and experimental results [3] .

The paper begins with a review section describing the four image metrics we used, and containing a brief description of their main properties as well as guide lines for computational algorithms. The next section presents comments regarding the experimental database and results. The following section deals with the relationship found between the experimental data and the metrics values, including quantitative correlation analysis. The experimental data and the image metric products for each of the tested scenes are shown in an appendix.

3. IMAGE METRICS

Image metrics, in general, are analytical or statistical procedures, designed to describe image properties quantitatively. An image metric can be either of global, local or combined global / local orientation. A global metric is a metric applied to the entire image, and returns a quantitative measure that represents an image property. Image properties presented by global metrics are, for example, image luminance, image intensities distribution and image clutter level. A local metric describes a property of a specific image area and is usually used to determine the target distinctness from its local surrounding. A well-known local image metric is the target to background contrast. The combined global / local metric integrates global and local image measures into one metric. The *SCR* - Signal to Clutter metric is an example of a combined global / local image metric. In this work we used four image metrics that were originally designed to evaluate human target detection performance of infrared imagery. These metrics are the global edge based clutter metric (*POE*), the local target to background distinctness metrics (*DOYLE*) and (*TARGET*) and the combined global / local texture metric (*ICOM*). As a direct consequence of the different perception processes of infrared and visible imagery by humans, we noticed degradation in the global metrics performance and improvement in the local metrics performance. Some of the reasons for that are:

1. The human visual system, usually being more experienced with visible imagery, is not easily attracted by natural changes in the image intensity even if the changes are sharp as is in the case with infrared imagery. Visible imagery, being characteristically more homogenous, we expect degradation in the global edge based clutter metric performance.
2. In most natural, military oriented, infrared images, the target to background distinctness is not a significant factor that influences the observer target detection performance. The reason for that is the fact that the target edges and inner details tend to smear out and the target looses its appearance of a man made object. Moreover, temperature

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

differences are not always significant to produce high local contrast levels.

These facts are supported by psychophysical experiments[2], which show that detection ability is dominated mostly by the global clutter level. On the other hand, in the case of visible imagery, the target color and sharp well-defined edges cause the target to background contrast to dominate the human perception process, and therefore one should expect to find good agreement between the local image metrics values and target detection performance.

As for the combined global / local metric, its performance depends on the image content, which determines the predominant component (global or local).

3.1 Probability of Edge Metric (*POE*)

The *POE* metric is a global image metric, originally designed to quantify infrared image complexity or clutter level. The metric is designed to imitate the human visual system, based on the assumption that the human eye fixates on image edges.

The technical details of the *POE* metric are as follows:

1. A *SOBEL* or similar spatial filter is used to enhance the image edges.
2. The output image intensity values are normalized to values between 0-255.
3. The resulting image is divided into N blocks of twice the apparent size of the typical target in each dimension.
4. The number of points that pass a predefined threshold T in each block i is counted and marked as the $POE_{i,T}$ value of the current block. The threshold T was chosen empirically to be 0.7 of the average pixel value of the original image in each process block.

The total image *POE* is defined as follows

$$POE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} POE_{i,T}^2} \qquad (1)$$

The *POE* metric was extensively tested with infrared imagery[2,4-6], yielding good results in quantifying image clutter levels and predicting human target detection probabilities. As mentioned above, for images taken in the visible spectral region, it is expected to be less efficient.

3.2 Local Distinctness Metric (*DOYLE*)

The *DOYLE* local metric is based on the target and its local background intensity values distribution. It is designed to measure the differences between means and variances of the target and the local background pixel values. The basic form of the metric is as follows

$$DOYLE = \sqrt{(\mu_t - \mu_b)^2 + K * (\sigma_t - \sigma_b)^2} \qquad (2)$$

where,

μ_t , μ_b , σ_t , σ_b - target and background means and variances.

K - weighting coefficient.

The *DOYLE* local metric was also tested extensively with infrared imagery[2,7] and produced acceptable results.

3.3 Local Target Complexity Metric (*TARGET*)

The *TARGET* local metric is based on the assumption that a target is more easily detected if its contrast to the background is high, if its size is big and if it has well-defined edge points relative to the interior points. While the first two conditions are trivial, the meaning of the third is that for easily detected targets the only noticed edge points are those which define the target contour. A quantitative measure for this property is defined as the target complexity (*TC*) and is equal to the area between the edge enhanced target image cumulative distribution function and a uniform cumulative distribution function as follows

$$TC = \frac{1}{L}\sum_{i=0}^{L-1} \left| S_N(i) - P(i) \right| \qquad (3)$$

where,

TC – target complexity.

L - No. of gray levels.

S_N - target cumulative distribution function.

P - uniform cumulative distribution function.

The expression above is also known as the Kolmogorov - Smirnov test.

Adding the effect of the target to background contrast and the target size (*TS*) yields the *TARGET* metric as follows

$$TARGET = TS\left(contrast + \alpha TC\right) \qquad (4)$$

where α is a fitting parameter (α=0.04 in our case).

Several works[4,5] are showing detailed explanation about the metric, as well as experimental results obtained with infrared imagery.

3.4 Combined Global / Local Texture Metric (*ICOM*)

The *ICOM* textural metric is based on the Markov co-occurrence matrix, which contains information about both the intensity values distribution and the possible transitions among neighbor pixels in the examined image area.

The co-occurrence matrix is a square (NxN) matrix where N is the number of possible pixel intensity values that occur in the image. Each image pixel contributes to the co-occurrence matrix according to its neighbor pixels intensities distribution. For the *ICOM* metric, the co-occurrence matrix is used twice. First, the textures of image areas in the size of the target are examined, looking for those areas that contain a target-like texture. Secondly, the textures of image areas of the size of the target and its local background are examined, looking for those areas that differ from the target size area texture. An intelligent combination of the two examinations can specify a target-like area, which is distinct from its local background. Such an area is probably either the real target, or an area very similar to it, attracting the human eye, and causing the human observer to classify it as a real target. The technical details for calculating the *ICOM* metric are as follows:

1. The target co-occurrence matrix – Ct is calculated in a square window containing only target pixels.
2. A square window of a size a bit smaller than the target size is stepped over the image. For each step the co-occurrence matrix of the area captured by the window – Cw is calculated.

3. The expression $x = \sum (Ct - Cw)^2$ is calculated for each step, measuring the difference between the target texture and the examined window texture.
4. A square window of a size twice the apparent target size is stepped over the image. For each step the co-occurrence matrix of the area captured by the window – Aw is calculated.
5. The expression $y = \sum (Cw - Aw)^2$ is calculated for each step, measuring the difference between the target size area texture and the examined window texture. We denoted this measure as *LOCAL ICOM.*
6. The empirical expression $z = (1-x)^4 * y^{\frac{1}{4}}$ combining the "target-like" term (x) with the "target distinctness" term (y) for each step, defines the *ICOM* image of the original scene.
7. All pixels in the *ICOM* image exceeding a predefined threshold are counted, and the result is defined as the *ICOM* value of the original scene. The larger is this value, more target like eye attracting areas are present in the image.

Further details as well as experimental results obtained using this metric are available in previous works[8,9].

4. COMMENTS ON THE EXPERIMENTAL DATABASE AND RESULTS

The experimental imagery database[3] includes 44, 4096*6144 pixels color images, each containing one of possibly nine military vehicles. The 44 images denoted img0001 to img0044 were divided into 3 groups. The first group contains 18 images, where the embedded target area is larger than 5000 pixels. Images indices in this group are 5, 9-10, 12-14, 18-20, 28, 31, 34-38, 40-41. The second group contains 10 images, where the extent of the target in either direction is less than 32 pixels. Images indices in this group are 4, 6, 11, 16, 23-24, 29, 32-33, 39. The third group contains the remaining 16 images.

The *DOYLE*, *TARGET* and *POE* metrics were implemented to the images of the second and the third groups, and the *ICOM* metric was implemented only to the images of the third group. The reasons for that are:

1. We believe that the very high detection performance (Pd>=0.94 for 17 images of 18) obtained for the images of the first group was dominated solely by the target size, regardless of the target distinctness or the image texture and clutter level.
2. Large computational efforts needed for the evaluation of the *ICOM* metric, forced us to degrade the image resolution. As a consequence, the targets embedded in the images of the second group appeared very narrow causing the algorithm to be less effective.

Another issue we considered regarding the imagery database is the contribution of color to detection performance. We conducted a simple, reduced scale experiment, by presenting the same images in color and in black and white to a group of observers, and asking them whether the color information helps them to detect a target, which they could not detect in the black and white image. The results of this simple experiment were almost absolute – the color information did not play any significant role in determining probability of detection. Based on this result we eliminated color from all the images, and implemented the metrics in the same way as done for the black and white infrared imagery for which they were originally designed.

5. ANALYSIS OF THE EXPERIMENTAL RESULTS AND METRICS VALUES

In order to analyze the agreement between experimental results and image metrics values, we plotted the values of the target detection experimental data as a function of the image metrics products, and calculated the correlation between them.

In ideal conditions we would expect the occurrence of a linear dependence, meaning, for example, that high detection probabilities will be associated with high *DOYLE* values and low detection probabilities will be associated with low *DOYLE* values. However, in practice we expect this dependence to be more ambiguous and to follow fuzzy-type logical rules.

The experimental data contains[3], for each tested scene, the number of correct, missed and false detections. Based on these results we calculated the probability of detection *(Pd)*, defined for every image as the ratio between the number of correct detections made by 62 observers, and the maximum possible number of correct detections that can be made by these observers. Another experimental measure we calculated for this analysis is the *Search Rate* defined as 1/ (Search Time).

Examples of the relationship between the experimental measures and the image metrics values are shown in figures 1-6.

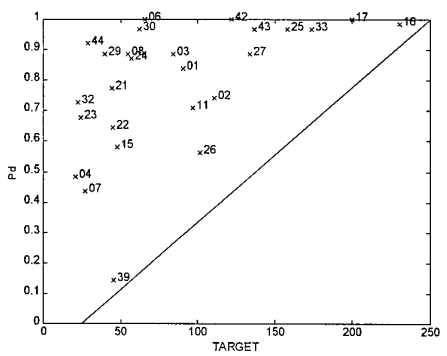


Figure 1. Pd = f(TARGET)

From figures 1 and 2 one can learn that high values of *DOYLE* and *TARGET* local metrics are, as expected, associated with high *Pd* values. Figures 3, 4.a and 4.b show the dependence of *Pd* upon *POE* values.

Generally speaking, it appears (fig. 3) that the global clutter level measured by *POE* is not a significant factor regarding detection performance. However, if the local target to background distinctness measured, for example, by the *DOYLE* metric is low, the image global clutter level becomes significant and dominates the detection probability (fig. 4.a). Moreover, if the *DOYLE* metric produces moderate to high values, the global clutter level has no significant role in determining the detection probability (fig. 4.b).

As for the results presented in figures 5 and 6, it appears that the scenes, which produce high *ICOM* are associated, as expected, with low value of search rate, and that the local part of the metric defined as *Local ICOM* is the dominant factor of this result.

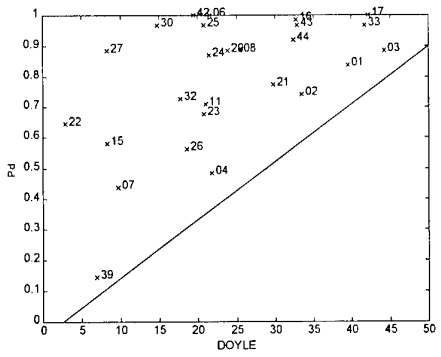


Figure 2. Pd = f(DOYLE)

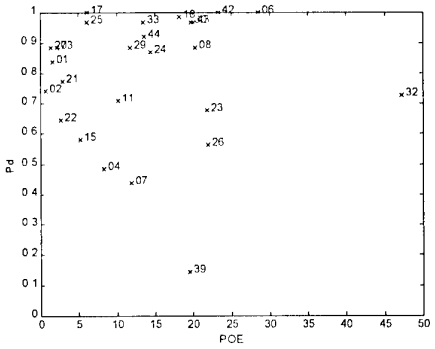


Figure 3. Pd = f(POE)

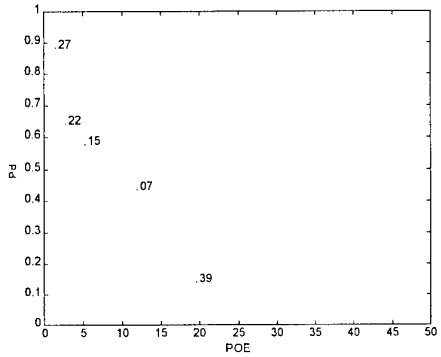


Figure 4.a Pd = f(POE $_{\text{LOW DOYLE}}$)

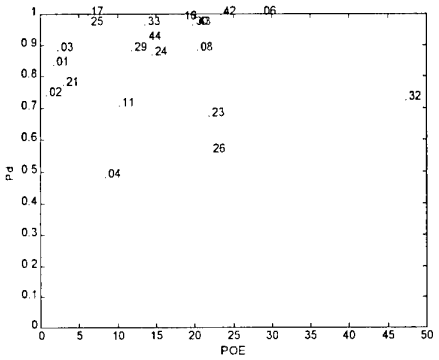


Figure 4.b Pd = f(POE $_{\text{MODERATE - HIGH DOYLE}}$)

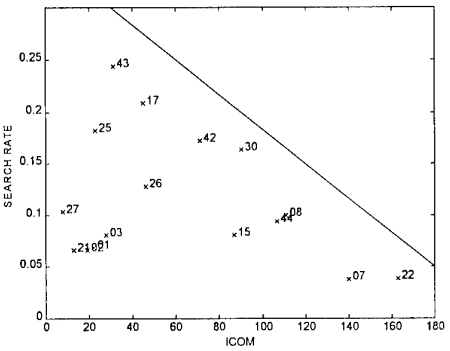


Figure 5. Search Rate = f(ICOM)

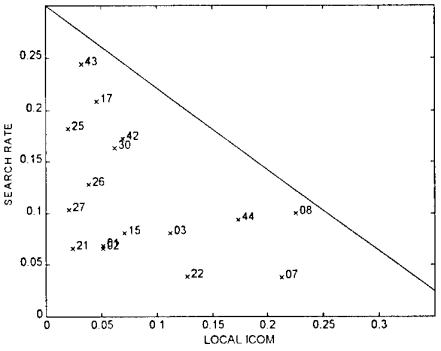


Figure 6. Search Rate = f(Local ICOM)

According to these figures, it also appears that the *DOYLE* and *TARGET* metrics can predict very well whether the embedded target will be easily detected, but they produce ambiguous results regarding low observable targets. The opposite occurs with *ICOM* and *Local ICOM*. This fact can be used to formulate empirical, fuzzy-type, classification rules. The fuzzy rules can lean on classification thresholds determined, for example, by the sloping line which divides the plane generated by the experimental results and the image metrics values into two regions. In order to quantify the results presented in figures 1-6, we used correlation analysis. The correlation (ρ) between two vectors A and B is defined as

$$\rho = \frac{COV(A,B)}{\sigma_A \sigma_B} = \frac{\sigma_{AB}}{\sigma_A \sigma_B} \qquad -1 \le \rho \le 1 \qquad (5)$$

where,

σ_A, σ_B - standard deviation of vectors *A* and *B* respectively.

COV(A,B) – covariance of *A* and *B*.

In our application, the vector A contains one of the performance measures (*Search Rate* or *Pd*) as calculated for all the evaluated scenes, while vector B contains one of the image metrics values (*POE* or *DOYLE* or *TARGET or ICOM*) as calculated for the same scenes.

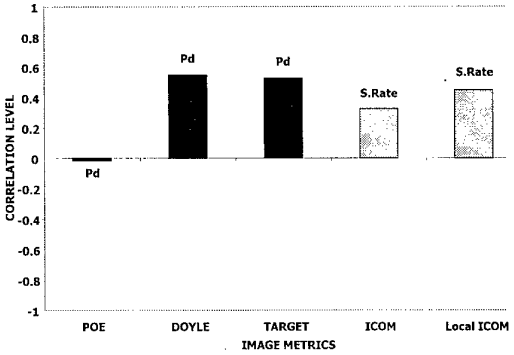


Figure 7. Correlation Values

For example, in order to calculate the correlation between the probability of detection performance factor, and the *ICOM* image metric values, we have to calculate the correlation between the two vectors *A* and *B* defined as

A= [Pd(Img1); Pd(Img2) ;...; Pd(Img7) ;...; Pd(Img44)]

and

B=[ICOM(Img1); ICOM(Img2) ;...; ICOM(Img7) ;...; ICOM(Img44)]

Representative results obtained from this analysis are shown in figure 7.

From the above figure one can learn about the relatively high correlation values obtained between both *DOYLE* and *TARGET* local metrics and *Pd*. It also appears that there is a moderate level of correlation between both *ICOM* and Local *ICOM* metrics and the *Search Rate*. Finally, there is almost no correlation between *POE* and *Pd*.

Just for the comparison, the correlation levels obtained with infrared imagery are 0.17 between the *DOYLE* values and *Pd*, 0.38 between the *POE* values and *Pd* and 0.67 between the *ICOM* values and *Search Rate* values.

6. CONCLUSIONS

The agreement between the TNO experimental results and various image metrics products was tested and evaluated in this paper. The analysis containing dependence and correlation measures is done after excluding images containing very large size targets (for evaluation of all the metrics), or very narrow extent targets (for evaluation of the *ICOM* metric). This is done because the experimental performance associated with these images is irrelevant in the context of the analysis methodology. The paper presents four image metrics, originally designed to evaluate detection performance of infrared imagery. These metrics have local, global and combined global / local orientation. Applying these metrics to the image database yields relatively high correlation values between the experimental results and the local metrics (*DOYLE* and *TARGET*) values. It also yields moderate correlation levels between the *ICOM* global / local texture metric values and the experimental results and very low correlation level between the *POE* global clutter metric values and the experimental results. These results emphasize that target to background distinctness is a significant perceptual cue regarding target detection in natural imagery taken in the visible region of the spectrum. Nevertheless, if the scene contains a low distinctness target, the global clutter level becomes relevant and determines the detection performance. These conclusions can be used to define fuzzy-type classification rules and to set further research directions.

REFERENCES

1. G.Aviram, S.R.Rotman, "Human Detection Performance of Targets Embedded in Infrared Images – The Effect of Image Enhancement".

   To be published in *Optical Engineering.*

2. G.Aviram, S.R.Rotman, "Evaluation of Human Detection Performance of Targets Embedded in Natural and Enhanced Infrared Images Using Image Metrics".

   Submitted to *Optical Engineering.*

3. A.Toet, P.Bijl, F.L.Kooi, J.M.Valeton, "Image Data Set for Testing Search and Detection Models", *TNO-Human Factors Research Institute*, report TM-97-A036, (Apr. 1997).

4. S.R.Rotman, G.Tidhar, M.L.Kowalczyk, "Clutter Metrics for Target Detection Systems", *IEEE Trans. Aerospace Electron. Syst.*, Vol. 30, No. 1, pp. 81-91, (Jan. 1994).

5. G.Tidhar, G.Reiter, Z.Avital, Y.Hadar, S.R.Rotman, V.George, M.L.Kowalczyk, "Modeling Human Search and Target Acquisition Performance: IV. Detection Probability in The Cluttered Environment", *Optical Engineering*, Vol. 33, No. 3, pp. 801 - 808, (Mar. 1994).

6. S.R.Rotman, M.L.Kowalczyk, V.George, "Modeling Human Search and Target Acquisition Performance: fixation-point analysis", *Optical Engineering*, Vol. 33, No. 11, pp. 3803 - 3809, (Nov. 1994).

7. A.C.Copeland, M.M.Trivedi, J.R.McManamey, "Evaluation of Image Metrics for Target Discrimination Using Psychophysical Experiments", *Optical Engineering*, Vol. 35, No. 6, pp. 1714 - 1722, (Jun. 1996)..

8. S.R.Rotman, D.Hsu, A.Cohen, D.Shamay, M.L.Kowalczyk, "Textural Metrics for Clutter Affecting Human Target Acquisition", *Infrared Physics & Technology,* Vol. 37, pp. 667-674, (1996).

9. G.Aviram, S.R.Rotman, "Evaluating Human Detection Performance of Targets and False Alarms, in Natural and Enhanced Infrared Images, Using an Improved COM (ICOM) Textural Metric.

   To be submitted to *Optical Engineering.*

APPENDIX

IMAGE	Pd	Search Rate	POE	DOYLE	TARGET	ICOM
1	0.8387	0.0685	1.6434	39.49	90.487	22
2	0.7419	0.0658	0.7298	33.5414	110.87	19
3	0.8871	0.0806	2.1878	44.2175	84.203	28
4	0.4839	0.0336	8.321	21.8674	20.979	Narrow Extent
5	Large Size Target					
6	1	0.1562	28.4679	21.5388	65.968	Narrow Extent
7	0.4355	0.0375	11.9183	9.7424	27.069	140
8	0.8871	0.1	20.3326	25.5205	54.583	111
9	Large Size Target					
10	Large Size Target					
11	0.7097	0.084	10.1294	20.9652	97.135	Narrow Extent
12	Large Size Target					
13	Large Size Target					
14	Large Size Target					
15	0.5806	0.0806	5.1394	8.313	47.942	87
16	0.9839	0.4	18.2353	32.7611	230.36	Narrow Extent
17	1	0.2083	6.117	42.1258	200	45
18	Large Size Target					
19	Large Size Target					
20	Large Size Target					
21	0.7742	0.0662	2.9254	29.8335	44.352	13
22	0.6452	0.0391	2.6516	2.8604	45.138	163
23	0.6774	0.0826	21.7671	20.7361	24.52	Narrow Extent
24	0.871	0.125	14.3901	21.4922	57.368	Narrow Extent
25	0.9677	0.1818	6.0804	20.8375	158.08	23
26	0.5645	0.1282	21.8908	18.6341	101.62	46
27	0.8871	0.1042	1.3562	8.281	133.84	8
28	Large Size Target					
29	0.8871	0.1333	11.8172	23.8755	39.795	Narrow Extent
30	0.9677	0.1639	19.7747	14.7332	62.26	90
31	Large Size Target					
32	0.7258	0.0813	47.2418	17.8229	22.484	Narrow Extent
33	0.9677	0.1852	13.5236	41.7339	173.67	Narrow Extent
34	Large Size Target					
35	Large Size Target					
36	Large Size Target					
37	Large Size Target					
38	Large Size Target					
39	0.1452	0.0287	19.5208	6.9069	45.353	Narrow Extent
40	Large Size Target					
41	Large Size Target					
42	1	0.1724	23.3694	19.4887	121.89	71
43	0.9677	0.2439	20.0798	32.9665	136.64	31
44	0.9194	0.0943	13.6477	32.4026	28.725	107

Table 1. Experimental data and image metrics products of the TNO database.

IMAGE BASED CONTRAST-TO-CLUTTER MODELING OF DETECTION

David L. Wilson
US Army CECOM NVESD
10221 Burbeck Rd., STE 430
Ft. Belvoir, VA 22060-5806
Phone: (+) 703 704 2106
Email: dwilson@nvl.army.mil

1. SUMMARY

Using image-based metrics, contrast-to-clutter modeling is applied to the Search-2 visible image set and perception experiment data. To calculate the contrast metric, a new image is generated from the original image by replacing the target with an "expected background" using the local background surrounding the target and the natural horizontal correlation present in most surface-to-surface scenes. The contrast metric is obtained from the difference of this new image and the original image. Via a simple mathematical formula, the ratio of the contrast measure to a clutter metric is used to predict performance.

Keywords: clutter, line PSS, contrast-to-clutter, power spectrum signature, PSS, probability of detection, RSS

2. INTRODUCTION

Models to predict probability of detection by human observers viewing images vary from contrast and resolution methodology, as in Johnson criteria for predicting IR FLIR performance, to complex visual models that attempt to consider the target and background spectral natures and the effects of those on the human detection process. Current modeling tends to have less than the desired accuracy in predicting probability of detection.

The work presented here originates from metrics used experimentally at NVESD for predicting IR FLIR performance in real imagery with cluttered backgrounds. The metrics used here were introduced at the 1997 SPIE Meeting (Ref. 1) and the 1998 Army Science Conference (Ref. 2). These metrics are combined together for the first time against a visible data set and perception data.

As the metrics and earlier modeling were against IR instead of visible imagery, the results here required the metrics to be evaluated against grayscale renditions of the original color images. It is possible that color equivalents for the metrics used might be found with further experimentation. However, for most of the images in the Search-2 data set, it is suspected, and somewhat confirmed by the following analysis, that color may not be a major factor in this image set.

A second caveat in this analysis is that the data set is relatively small with most targets having relatively high probability of detection. A broader range of probability of detection would be more desirable for a robust evaluation.

Finally, as the originally provided data set had errors in the ranges for images 35, 38 and 43, these images were excluded from the analysis as time did not permit using the corrected ranges that were later provided for those three images.

3. IMAGE METRICS

Image metrics used here fall into four categories: contrast metrics, clutter metrics, size metrics and shape metrics. The following will introduce the objective contrast, clutter and size metrics used in the modeling.

3.1. Contrast Metrics

In general terms, a contrast metric is a metric that measures the intensity difference between a target and its local background. Such metrics may be in gray levels, light intensity levels, or in the case of IR FLIR imagery in temperature units. Usually, the local background is taken to be a box with dimensions (width and height) the square root of 2 multiplied by the dimensions (maximum width height) of the target. In the case of a rectangular target, this gives the local background the same area as that of the target.

The simplest contrast metric is the difference between the target and background means:

$$\Delta\mu = \left|\mu_{tgt} - \mu_{bkg}\right| \quad \text{(Eq .1)}$$

where μ_{tgt} is the target mean intensity and μ_{bkg} is the background mean intensity.

The difficulty with the above metric is that it does not consider internal structure of the target and background. The target and background may have the same means but the target may be detectable due to its internal structure. One of the most commonly used contrast measures that attempts to somewhat correct this problem is the RSS (Root Sum-of-Squares). The RSS is given by:

$$RSS = \left[\frac{1}{POT}\sum_{pixel(i,j)\in tgt}(t_{i,j} - \mu_{bkg})^2\right]^{1/2} \quad \text{(Eq. 2)}$$

where $t_{i,j}$ is the intensity of the pixel (i,j) and POT is the number of pixels on target. The RSS can be calculated readily from the target and background means and the target standard deviation by the following alternative formula:

$$RSS = \left[\left(\mu_{tgt} - \mu_{bkg}\right)^2 + \sigma_{tgt}^2\right]^{1/2} \quad \text{(Eq. 3)}$$

where σ_{tgt} indicates the standard deviation of the target.

A different contrast metric that has been proposed (Ref. 1) is an implementation of the PSS (Power Spectrum Signature). This requires us to define an "expected background" image for what we might expect to see if the target were not present.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

Naturally, such an image cannot be precisely exactly defined. For the moment, assume we have such an expected background image; we may then define the PSS as:

$$PSS = \left[\frac{1}{POT} \sum_{i,j} \left(t_{i,j} - b_{i,j} \right)^2 \right]^{1/2} \quad \text{(Eq. 4)}$$

where the summation is over all pixels which are different in the original image and expected background image, t_{pix} indicates the intensity of target pixels and b_{pix} indicates the intensity of expected background pixels.

The problem is to now define a usable concept of expected background. Note that this is not the same as the actual background. For example, the actual background might contain a hot or bright rock. We would not expect to see such. The expected background should be one that does not draw the attention of the observer. There are various possible implementations for an expected background. For example, one might replace the target by the mean of the local background. If one actually does this, one quickly discovers that the flat intensity that results is far from expected and readily draws one's attention.

The implementation of the expected background that is used here is based on the fact that real images of surface-to-surface scenes tend to have high horizontal correlation. This is probably due to two primary factors. The first is basic geometry. Even a solid circle patch on the ground at a distance horizontally will appear to be an ellipse with the major axis in the horizontal direction. Another factor to consider is that local to the target, the horizontal will tend to be at the same range and suspect to the same propagation effects as well as contain the same vegetation. There are exceptions to this: for example, long exposed tree trunks would give a strong vertical correlation in that part of an image. In general though, the horizontal correlation might be expected.

The above leads to the concept of the (horizontal) line expected background and PSS. We define the expected ground intensity at target pixel locations to be:

$$b_{i,j} = \left(1 - i'/n(j) \right) \mu_L + \frac{i'}{n(j)} \mu_R \quad \text{(Eq. 5)}$$

where i' is the distance along the horizontal from the left edge of the target to pixel (i, j), n(j) is the distance along the j^{th} horizontal inside the target, μ_L is the mean intensity on the horizontal in the local background to the left of the target and μ_R is the mean intensity on the horizontal in the local background to the right of the target (see Fig. 1). In concept, the line PSS is a linear horizontal interpolation between the mean local left intensity and the mean local right intensity.

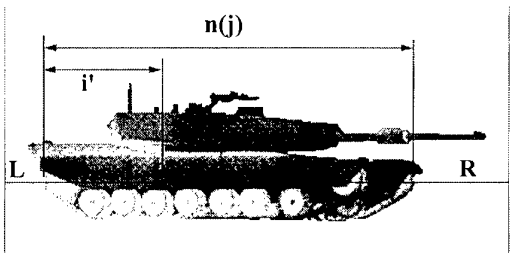


Figure 1: Calculating Line PSS

The image formed with the line PSS expected background often works surprisingly well. For example, see Fig. 2a, Fig. 2b, Fig. 3a and Fig. 3b.

Figure 2a: Crop of target in image 10

Figure 2b: Line PSS expected background of Fig. 2a

As expected, if the target is too close (large) and surrounded by clutter, the expected background often appears too flat or to have unusual horizontal streaks. Also, if there is a high contrast clutter object horizontally in the local background of the target, this will cause a conspicuous horizontal streak in the expected background image. In such cases, the line PSS expected background methodology needs to be modified; but this has not been done in the analysis that follows.

3.2. Clutter Metrics

The simplest clutter metric is the standard deviation of the image or the standard deviation of the local background σ_{bkg}. In practice, this does not seem to work very well.

Dr. Silk at the Institute for Defense Analysis has proposed another clutter metric (Ref. 3). It is a modification of the Schmieder-Weathersby clutter metric (Ref. 4) and bares some similarity to the form of the line PSS. As we will later form the ratio of the contrast metric to the clutter metric, similarity in form is a desirable property.

Figure 3a: Crop of target in image 33

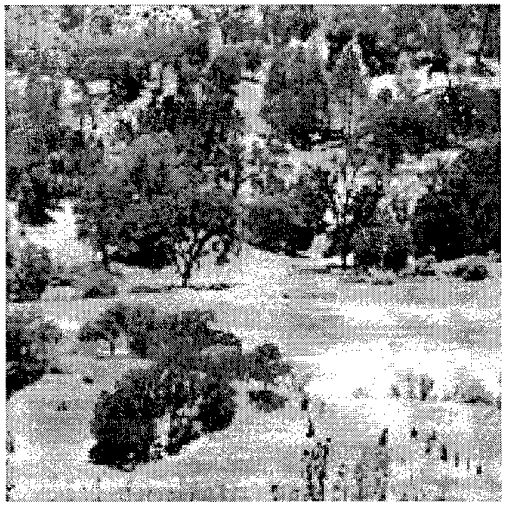


Figure 3b: Line PSS expected background of Fig. 3a

The implementation of this clutter metric may be visualized as a calculation obtained by convolving a square box centered on an image pixel through the image. There are two ways to handle difficulties encountered at the image border. The first is to pad the image by reflecting the image, or padding with the mean or some other constant. The second is to only convolve the box to positions that keep the box interior to the image. As little difference it obtained between these two choices and the second is easier to implement, it is the one used in the following analysis. The formula for the implemented clutter metric, denoted CM, may be expressed as:

$$CM = \left[\frac{\sum_i \sum_j \left(b_{i,j} - \overline{B}_{i,j} \right)^2}{N} \right]^{1/2} \qquad \text{(Eq. 6)}$$

where $b_{I,j}$ is the intensity of the (i, j) pixel, $B_{I,j}$ is the square box of pixels centered at (i, j) with the bar above it indicating the mean intensity of the box and N is the number of boxes convolved in the image.

As the target should not be considered cluttered, the clutter metric CM is actually calculated on the line PSS image having the expected background. This is even more important when multiple targets are present in the image, as not eliminating them from the calculation will cause too large a value for the clutter metric. Also, horizons should not be allowed in the calculation of CM, as they will dominate the calculation. One could similarly argue that other strong contrast, clearly recognized structures should also be eliminated from the CM calculation, as they probably do not act as clutter. But no procedure is currently proposed for doing this.

The size of the box has been subject to experimentation. Generally it has been set at 4 meters at the target range. This is larger than most targets. But should the box represent actual target size, expected target size or some other size? Clearly more work needs to be done concerning the box size; but in the past, using 4 meters at target range seems to have worked as well as any other choice and is used here.

Another issue is whether the clutter metric should be calculated over the entire image or over some region more local to the target. In the analysis that follows, CM was calculated over both the entire image and over a region subjectively local and similar to that near the target.

3.3. Target Size

The target size used in this analysis was the square root of the pixels on target. The software used to calculate the line PSS and CM required using a re-sized to 768x512 gray scale version of the images and re-segmentation of the targets. The square-root number of pixels on target (POT) from that segmentation was used as the target size.

3.4. Metric Values

Although the next section will model probability of detection using metrics described in this section, for completeness, a table of measured metrics is included here.

In Table 1, CM1 indicates the clutter metric calculated on the entire image while CM2 indicates the clutter metric calculated on a subjectively determined region containing the target that is typical to the area around the target. As one readily sees, CM is sensitive to the region chosen and that has caused some concern. But in truth, the difficulty is a lack of knowledge on what should really be considered clutter, rather than a problem with the calculation.

As mentioned earlier, images 35, 38 and 43 are excluded due to errors in the originally provided ranges that are used in determining the box sizes in the calculation of CM. Both the contrast and clutter metrics are measured in gray levels.

4. PROBABILITY OF DETECTION MODELING

Various combinations of the metrics might be considered for modeling probability of detection. Among these are contrast (such as RSS or line PSS) alone, size (such as square root of POT) alone, contrast times size and contrast times size divided by a clutter metric (such as the standard deviation of the local background or CM). This last predictor for probability of detection is loosely referred to as contrast-to-clutter. Of particular interest is the case of line PSS as the contrast metric and CM as the clutter metric. But other predictors will be also considered for comparison.

In each case, one generally expects the larger the above predictor, the larger the probability of detection. The natural goal is to find one that works "best" in general. "Best" can be defined in various, often conflicting, ways: least scatter, least

Table 1: Measured Metrics

Image	POT	RSS	Line PSS	CM1	CM2
1	38	38.95	32.76	6.8	8.53
2	60	32.99	34.66	5.88	8.13
3	42	43.71	41.42	6.17	7.58
4	14	19.07	29.58	7.28	9.66
5	707	60.82	68.54	17.04	21.43
6	24	23.03	30.09	10.64	12.85
7	13	17.7	24.95	7.44	9.02
8	22	30.7	32.98	8.29	9.9
9	899	35.33	40.31	14.66	16.59
10	79	43.74	48.24	9.37	12.13
11	43	21.79	26.52	7.77	10.79
12	452	44.91	61.04	16.81	19.47
13	152	25.36	36.13	10.57	11.04
14	185	35.34	38.74	11.51	14.72
15	16	21.17	25.07	5.54	6.75
16	77	33.57	39.79	7.53	9.32
17	80	43.86	52.47	12.89	11.81
18	193	28.29	51.13	7.88	15.56
19	146	20.67	19.45	13.84	14.65
20	203	50.92	58.62	12.48	13.16
21	15	41.56	44.98	8.54	10.31
22	18	15.69	29.65	7.52	8.3
23	14	30.05	47.25	9.07	11.19
24	25	27.19	25.73	10.68	6.96
25	37	38.93	49.34	10.3	10.14
26	37	24.62	22.37	8.03	9.76
27	19	36.38	49.51	6.93	8.7
28	80	36.96	29.84	11.75	13.38
29	17	36.67	37.64	8.74	9.91
30	34	30.4	44.38	9.9	12.76
31	245	39.84	39.78	12.79	20.42
32	7	32.89	35.27	8.75	10.33
33	47	48.96	44.8	11.24	14.41
34	1162	70.83	87.93	23.66	25.95
36	92	35.07	59.05	12.91	17.13
37	438	48.47	79.18	17.46	30.21
39	22	14.77	24.37	7.88	13.67
40	144	63.92	65.39	14.49	12.16
41	388	38.34	40.84	17.64	23.55
42	42	38.16	48.98	11.33	14.72
44	20	37.24	32.08	7.11	8.81

RMS error or largest correlation between measured and predicted probabilities.

For the modeling that follows, the equation

$$PD_{pred} = \frac{(X/X_{50})^E}{1+(X/X_{50})^E} \quad \text{(Eq. 7)}$$

is used to predict the probability of detection. In the equation, X represents the predictor or combination of metrics used in the prediction. Both E and X_{50} are determined by non-linear regression (the Levenberg-Marquadt least-squares method).

In the table below, r is the Pearson product-moment correlation, COD is the coefficient of determination (r^2) and r_s is the Spearman rank correlation. The larger the value of r, the better the correlation between the predicted and measured probabilities.

Table 2: Prediction Models

X	X_{50}	E	COD	r	r_s
RSS	17.2	3.63	0.65	0.81	0.58
RSS•√POT	71.2	2.12	0.71	0.84	0.77
RSS•√POT/σ_{bkg}	3.20	1.58	0.55	0.74	0.49
PSS	20.4	3.20	0.42	0.64	0.59
PSS•√POT	93.5	2.37	0.62	0.79	0.77
PSS•√POT/σ_{bkg}	4.45	1.70	0.52	0.72	0.58
PSS•√POT/CM1	9.76	1.91	0.42	0.65	0.65
PSS•√POT/CM2	9.68	2.72	0.63	0.79	0.70

Note that COD measures the relative amount of the measured variance accounted for by the model only if one assumes the residuals follow a normal distribution with constant variance. The Spearman rank correlation has no such requirement on the distribution and variance.

It is interesting that all the better models above may be working roughly equally well when one considers the sample size. Some of the models yielding the larger correlations above do not include a clutter metric. One would conjecture they would perhaps not do as well if there were a greater range in image clutter. Additionally, if the "gain" of the display is adjusted, the contrast-to-clutter metrics have the advantage that this effect should somewhat cancel. Within some reasonable range, assuming relative linearity between gray levels and screen brightness this seems reasonable.

Using CM2 instead of CM1 appeared to give a better correlation. Recall that CM2 is the clutter metric CM measured over a subjectively determined region containing the target that was judged similar to the area near the target, rather than using the entire image as in CM1. This region contained a tree line if the target was near one; otherwise, it did not. If there were trees near the target, the region contained as large a group of similar trees as could be selected. Different people might select different regions. One might hope to formalize this process. As was mentioned earlier, a better understanding, and probably also of measurement, of clutter is needed.

Figures 4a-b shows plots of PSS•√POT/CM2 versus measured probability of detection for the perception experiment data. The curve is the model prediction regression line. Figure 5a-b are similar plots of the case of RSS•√POT for comparison.

Although both the RSS•√POT based and PSS•√POT/CM2 based models appear to work nearly equally well, as noted earlier, the contrast-to-clutter model should in theory have advantages when there are large variations in clutter or changes in the display gain.

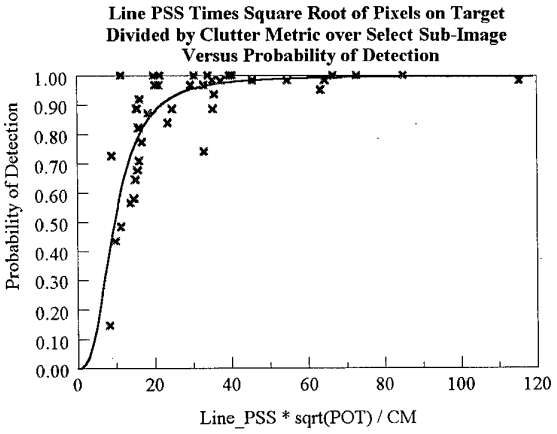


Figure 4a: PSS•√POT/CM2 vs. Probability of Detection

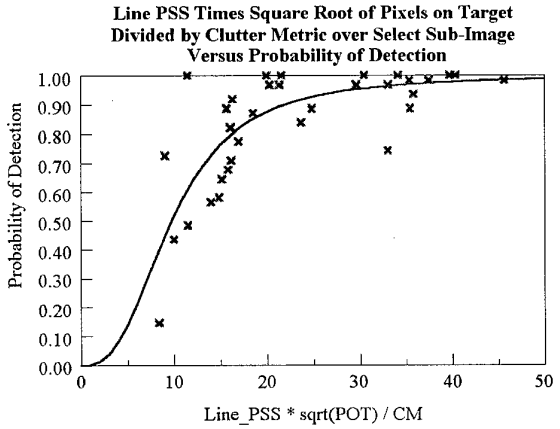


Figure 4b: PSS•√POT/CM2 vs. Probability of Detection enlargement of left side of Figure 4a.

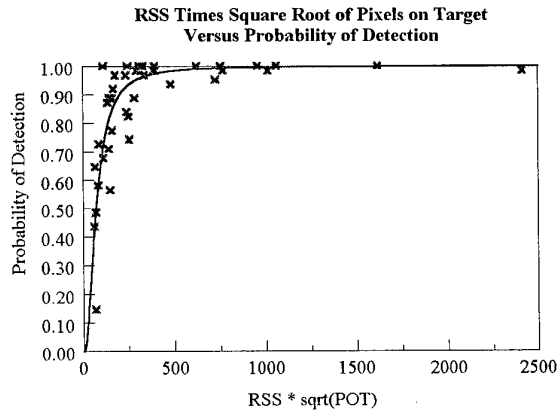


Figure 5a: RSS•√POT vs. Probability of Detection

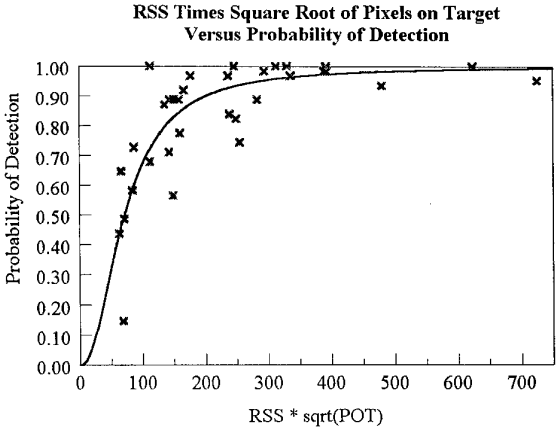


Figure 5b: RSS•√POT vs. Probability of Detection enlargement of left side of Figure 5a.

The outlier with probability of detection equal to 1.00 and least prediction metric in each case is image 6. This target casts a very conspicuous shadow that draws an observer's attention. Although the shadow was segmented with the target in the line PSS calculation, the shadow is probably a very strong cue to the observer.

5. CONCLUSION

The metrics used in this paper perform well enough against the Search-2 data to warrant further investigation in predicting visible probability of detection. The proposed contrast-to-clutter methodology has the promise of being somewhat self-calibrating. The studied metrics do not include color. Also, the metrics in the modeling presented do not include shape, although shape is also obviously important in target detection. Attempts to use a subjective shape measurement (Ref. 2) were unsuccessful with this data set. This may be due to the rather limited data set or the correlation between size and observer recognition of shape. The roles of contrast, size, shape and color are complicated by correlation between these factors. Additional research is needed that realizes this correlation.

6. REFERENCES

1. D'Agostino, J., Lawson, W., Wilson, D., "Concepts for search and detection model improvements", *Proceedings of SPIE*, Vol. 3063, pp.,14-22, 1997. Also available on the CD-ROM *Selected SPIE Papers on CD-ROM: Infared Technology 1988-1998,* Vol. 1, 1998.
2. D'Agostino, J. A., Wilson, D. L., Blecha, B. A., Biederman, I., Lawson, W. R., "The Role of Shape in Human Search & Detection", *21st Army Science Conference Proceedings* , 1998.
3. J. D. Silk, *Statistical Variance Analysis of Clutter Scenes and Applications to a Target Acquisition Test*, IDA Paper P-2950, November 1995.
4. Schmieder, D. E., Weathersby, M. R., "Detection Performance in Clutter with Variable Resolution", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 19, No. 4, p. 622, July 1983.

EFFICIENT METHODS FOR VALIDATING TARGET ACQUISITION MODELS

R. Hecker
IABG
Einsteinstrasse 20
85521 Ottobrunn
Germany
E-mail: hecker@iabg.de

1. SUMMARY

On the basis of target acquisition fundamentals the camouflage assessment model CAMAELEON is presented and especially demands, principles and methods for validating the model. By indirect varying the distance to a target using zoom techniques of telescopes effective methods for validating the model have been developed in the visual range as well as in the infrared range.

The paper presents the results of validation studies in the visual range and results of CAMAELEON model calculations with the SEARCH DATA image set made available by the TNO Human Factors Research Institute.

The results are discussed on the basis of the underlying principles of the CAMAELEON model and the SEARCH DATA evaluations especially of visual lobe.

Further investigations on the development of CAMAELEON are presented on the basis of the gathered experiences.

Keywords: Target acquisition, validation, detectability, detection, perception, visual lobe, camouflage, visual, infrared

2. INTRODUCTION

A shortcoming of many target acquisition models is their lack of validation. The main reasons for this lack may be the following:

- Target acquisition models in many cases are very complex with numerous parameters. To cover and/or control all these parameters the statistical sample size in validation field trials has to be very large for significant results.
- Military field trials with a lot of test persons and military target acquisition tasks are time and cost consuming. Because of restrictioned funds and time, field trials often do not result in a sufficient sample size for validating target acquisition models.

	Definition	Depends on	Characteristics
1. Detectability	ability to distinguish between object and background, decides, whether a certain object **can** be detected	size luminance contrast texture color shape primitives motion	„global" perception low level vision preattentive, without cognitive processes („automatically") figure ground separaration texture segregation low intra- and inter-individual variability
2. Detection	classification into objects and background (in a real world, e.g. military , natural environment), decides, whether a certain object **is** detected	in addition to 1.: varying weather conditions visual complexity of natural scenes search process, search area „briefing" of the observers attention, fatigue, training	takes place, if the object is detectable according to 1. **and** if the object is „fixated"
3. Recognition	classification of objects, e.g. into types (generic classification)	in addition to 1. and 2.: shape detection (general) knowledge of the observers	„specific" perception recognition of details for classification
4. Identification	classification within type into e.g. military individuals (specific classification)	in addition to 1. - 3.: (specific, e.g. military) knowledge of the observers	„specific" perception recognition of details for specific classification

Table 1. Target acquisition fundamentals

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

To get an answer on how to overcome these shortcomings Table 1 gives a survey of target acquisition fundamentals. Generally speaking from 1. (detectability) to 4. (identification) we find the following coherences:

a. Perception process: *in*creasing complexity
*in*creasing intra- and inter-individual variability

b. Modeling: *de*creasing knowledge about the perception process
*in*creasing number of parameters
*in*creasing complexity of the models

c. Validation: *in*creasing number of parameters to be controlled
*in*creasing interfering effects (weather conditions, learning, motivation, etc.)
*de*creasing statistical significance
*in*creasing necessary statistical sample size
*in*creasing demand of resources (cost and time)

So mainly two things should be done: *The target acquisition models should be reduced to the basics of the acquisition process* (as far as possible, depending on the object and use of the model), and *the validation field trials should be adapted to the question of the model.*

3. CONCEPT OF CAMAELEON AND VALIDATION

CAMAELEON is a computer model developed for the *assessment of camouflage* using digital image processing techniques based in part on the human visual system (Hecker, 1992).

As camouflage mainly depends on the similarity between an object and the nearby background, CAMAELEON is confined to the basics of acquisition in the sense of section 2. It aims at measuring the physiological *detectability* of an object against the nearby background by describing the similarity between object and background relating to first order statistic features like *contrast* and textural features like *local contrast (energy), local spatial frequency and local orientation.*

These local textural features are calculated from the output of several bandpass-filters which are similar to the filters constituted by the receptive fields of the neurons in the early stages of the human visual system.

For object and background separately the histograms of these local features and their overlaps can be calculated to obtain measures for similarity between object and background.

These similarity measures are combined in a heuristical detection model to calculate the detectability probability as a function of range and the detectability range.

Based on this concept of the CAMAELEON model the main principles of the field trials carried out to validate CAMAELEON are:

a. *Direct measurement of the detectability ranges of the objects.* The subjects know the position of the object. By varying the distance to the object they have to mark the distance from where the object just can no longer be discriminated from the background (or respectively only just the object can be discriminated from the background).

b. *Small objects.* This avoids large detectability ranges, reduces the duration of measurements, thus reducing interfering effects, especially atmospheric effects because of relatively constant surrounding conditions during the measurements. In addition no time consuming changing of objects and position of objects and the subjects were necessary.

c. *Indirect variation of the distance from the subjects to the object* by using zoom techniques of a special telescope and thus having *constant distances to the objects* with *no interfering atmospheric effects.* This further reduces the duration of measurements, increases the probability of constant surrounding conditions and permits tests of many different objects and backgrounds in a short time.

d. *Controlled variation of the parameters which do influence the detectability* ***and*** *are subject of the CAMAELEON model:* Size, contrast and texture.

e. *Controlling as far as possible the parameters which do influence the detectability, but are* ***not*** *subject of the CAMAELEON model:* Reduced variability of light conditions (see c.) and colors of object and background, constant shape of the object.

f. Parallel to the field tests *taking the images of the scenes* from the observer positions for further evaluation with the CAMAELEON model.

These principles were applied to field trials in the visual range as well as in the infrared range.

3.1. CAMAELEON Validation in the Visual Range

From Fig. 1 it can be seen, that for the field trials in the visual range only one observation point has been chosen with the measuring telescope and the camera. Five different target positions and according to this five different backgrounds have been chosen. So with 8 different targets of different size and texture in total 40 different scenarios for one trial session could be utilized. The distance of the observation point to the targets was 30 m.

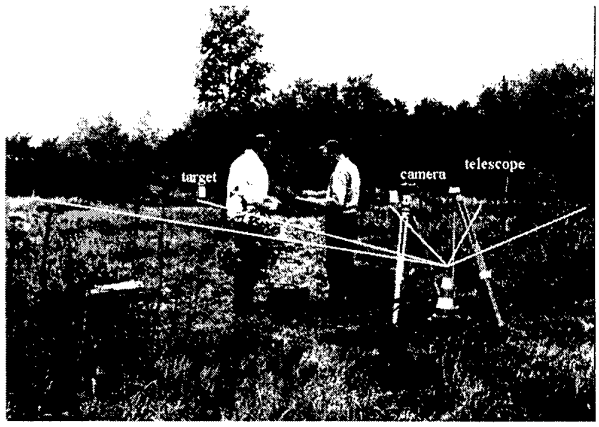


Fig. 1 Scene of the validation field trials in the visual range

Fig. 2 shows examples of chosen scenarios with different textures of object and background.

The measuring telescope from CARL ZEISS (Fig. 3) was used inverse and the subjects had to look through the object lense. Thus the measuring telescope had a reducing effect. To guarantee that the observers had a central view through the object lense, a tubus with a hole of 10 mm in diameter was attached in front of the object lense.

A specific scaled tuning of the adjustment control simulated a specific distance to the target. This scaling has been realized in a preceding study, so in the field trials *changing the distance to the targets was achieved by tuning the adjustment control.*

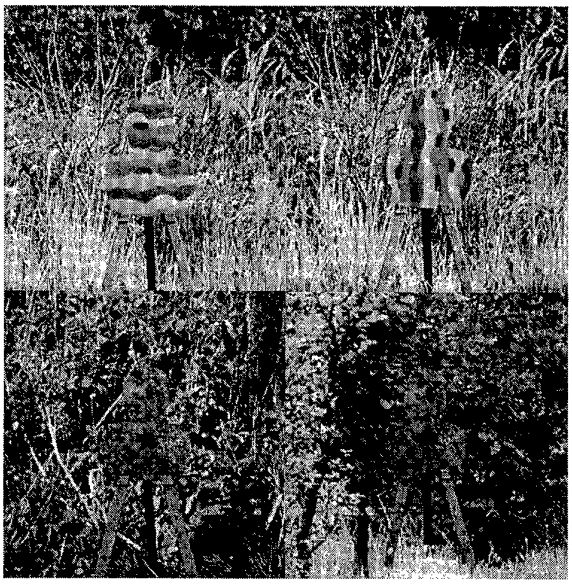


Fig. 2 Examples different textured objects and backgrounds

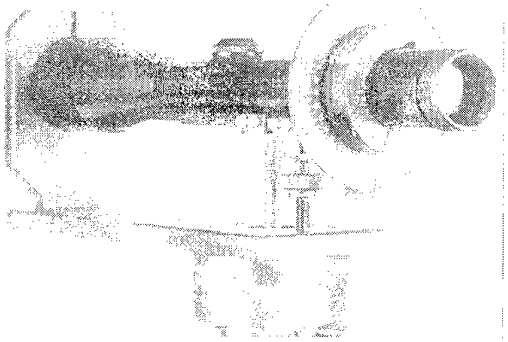


Fig. 3 Measuring Telescope with adjustment control

The apparent distance range which could be be tuned in the field trials was from 37.5 m to 600 m.

With the techniques described above a very short duration of measurements could be achieved, and within a one week trial it was possible to get a sample size of up to 120 for three to five observers.

In preceding studies it was found a very high consistency among observers, that is the inter-individual correlation was greater than 95%. So only few observers are necessary for the validation studies.

The calculation of the CAMAELON model weighting factors by correlation maximizing of measured and calculated detectability ranges has been done with a sample size of 120 assessed images.

As a result CAMAELEON showed $r = 0{,}81$ correlation (PEARSON) with another set of 120 assessed images, that is *r^2 = 66% of the variability of the measured detectability ranges could be attributed to the CAMAELEON model.*

3.2. CAMAELEON Validation in the Infrared Range

The same principles for validation as described in section 3. have been applied to the infrared range. Especially the distance to the objects has been kept constant. Instead of this varying distance has been simulated by changing the variable zoom of the used standard IR system TICM II (8_μ - 12_μ).

In the field trials only videos were taken of the scenes while zooming the scene within the whole zoom range. The videos can be evaluated later with subjects in a room with dusky illumination. The subjects have to stop the video, when the object just can no longer be discriminated from the background (or respectively only just the object can be discriminated from the background). From special marks in the IR-images the simulated distance to the object can be recalculated.

A major problem was to get temperature stabilized thermal textured IR-targets. This problem has been solved by using heatable and temperature stabilized boards, reflecting the radiance via aluminium targets to the observer. Texture has been created by paintings on the aluminium targets, thus varying the emissivity of the target surface.

Fig. 4 shows the setting up of the used equipment. To not disturb the measurements the heatable board has been "camouflaged" by IR effective nets in the direction of the observer.

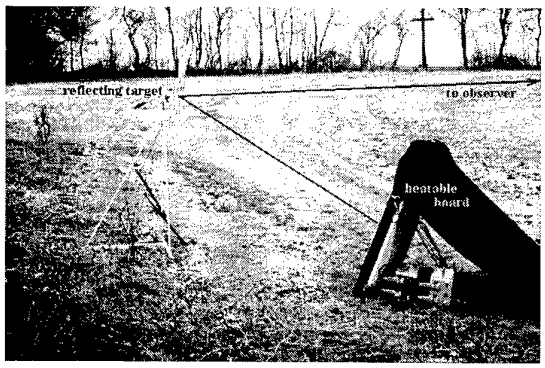


Fig. 4 Scene of the validation field trials in the IR range

In Fig. 5 an example of a IR-scene with a simple thermal textured object can be seen. On the bottom right parts of the covered heatable board can be seen.

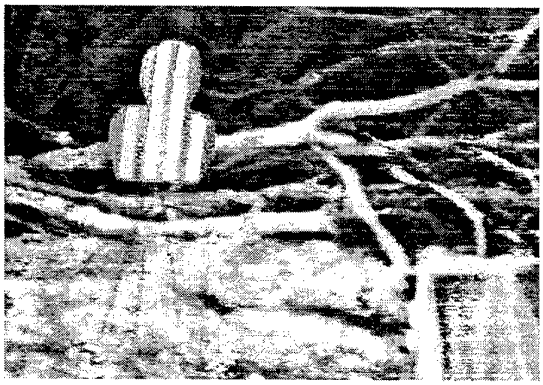


Fig. 5 IR-scene with simple textured object

Till now the CAMAELEON model doesn't contain an infrared sensor model, so in a first step only the correlation between measured and calculated detectability ranges can be calculated.

Because of small sample sizes till now and problems with getting calibrated data from the images evaluation has not

been finished yet, but the method itself seems to be very effective, although because of higher technical expenditure the number of evaluable scenes in a certain period of time is much less then in the visual range.

Analyzing the available data and the data from other IR studies suggest, that in many cases the thermal contrast between object and background is so high ("hot spots") , that according to the large detectability ranges textural features don't play a dominant role concerning detection, so CAMAELEON for the infrared range has to be adapted to this special situation.

4. CAMAELEON RESULTS ON THE SEARCH DATA

Two main aspects have to be considered when analyzing the Search Data (Toet et al., 1998) with CAMAELEON:

First CAMAELEON has been validated with standardized images, that is: taken from nearby, high resolution images of the objects, no atmospheric effects, while the images of the Search Data where taken from a wide variety of distances. This also results in a wide variety of object sizes and resolution of the objects in the screen situation which has been used for evaluation of the Search Data.

Second CAMAELEON tries to measure *detectability range* analyzing the nearby surround of the target, that is

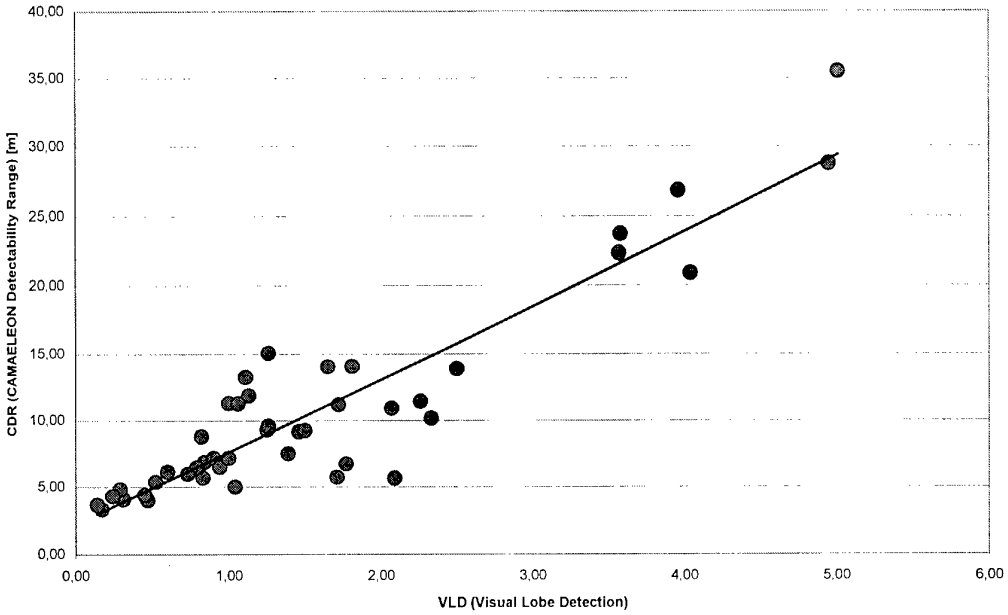


Fig. 6 CAMAELEON Detectability Range - Visual Lobe Detection

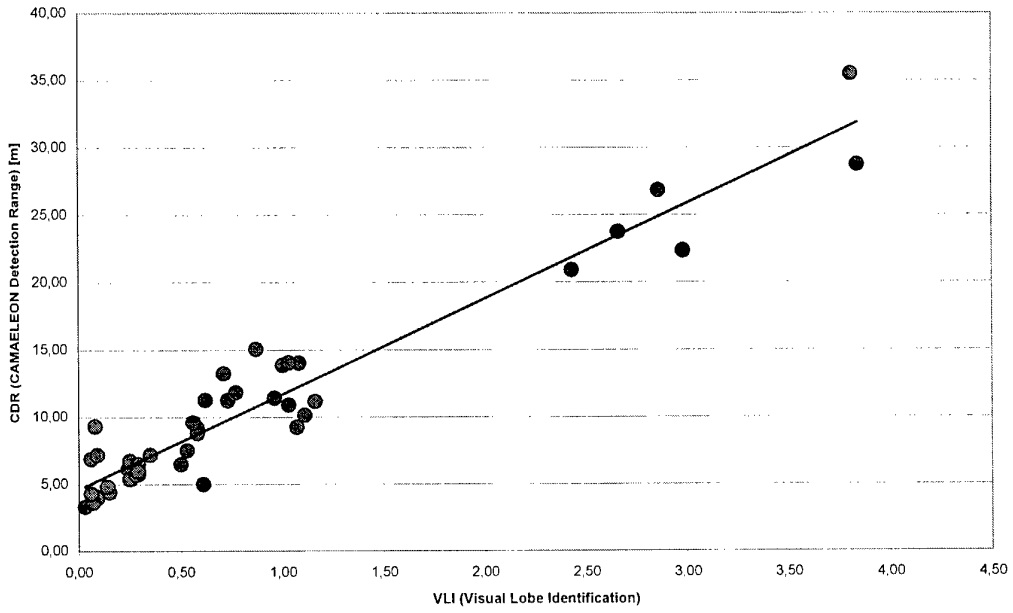


Fig. 7 CAMAELEON Detectability Range – Visual Lobe Identification

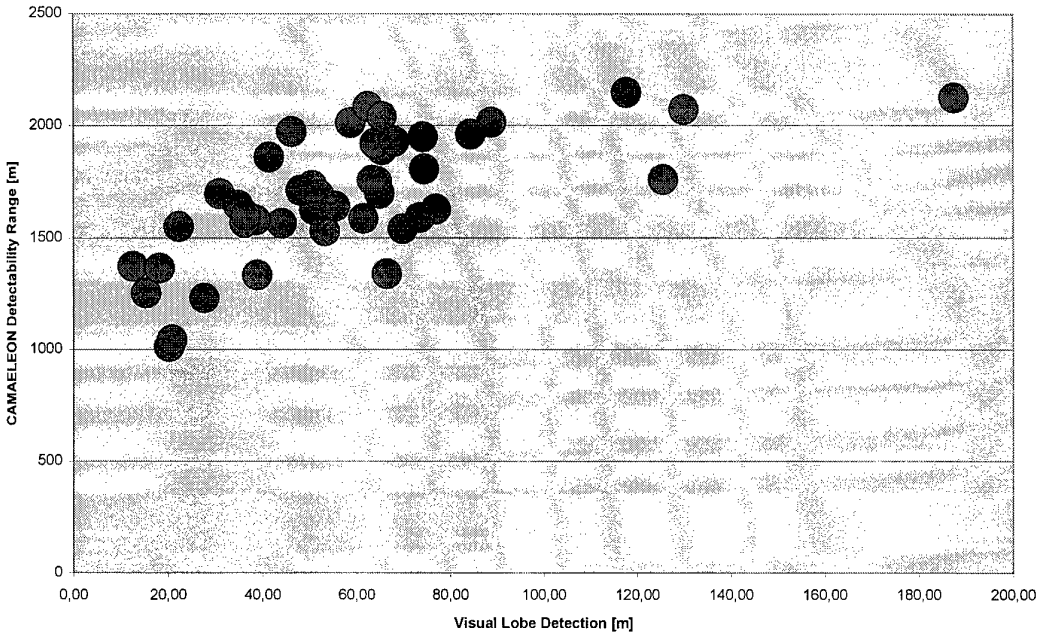


Fig. 8 CAMAELEON Detectability Range – Visual Lobe Detection („real word")

CAMAELEON uses *local metrics*, while the Search Data have been evaluated with *visual lobe* and *search time*, which also depend on the overall structural composition of the scene.

In the sense of section 2. the *Search Data field trials are not adapted to the question of the CAMAELEON model.*

So in general it is expected that the CAMAELEON results should be worse than those of (validated) models which involve *semi-local metrics* and /or *global conspicuity metrics* of the overall scene.

In particular it is expected, that the CAMAELEON results depend on the viewing distance of the Search Data images, that is the results should become better with decreasing target to the camera distances and thus decreasing the difference between nearby area around the target and entire scene.

As CAMAELEON doesn't include higher order processes as searching, the search time results of the Search Data have not been compared with CAMAELEON results.

4.1. Evaluation of the screen situation

In a first step the detectability ranges in the screen situation have been calculated, that is all targets had the same distance to the observer, the size of the targets was that of the size of

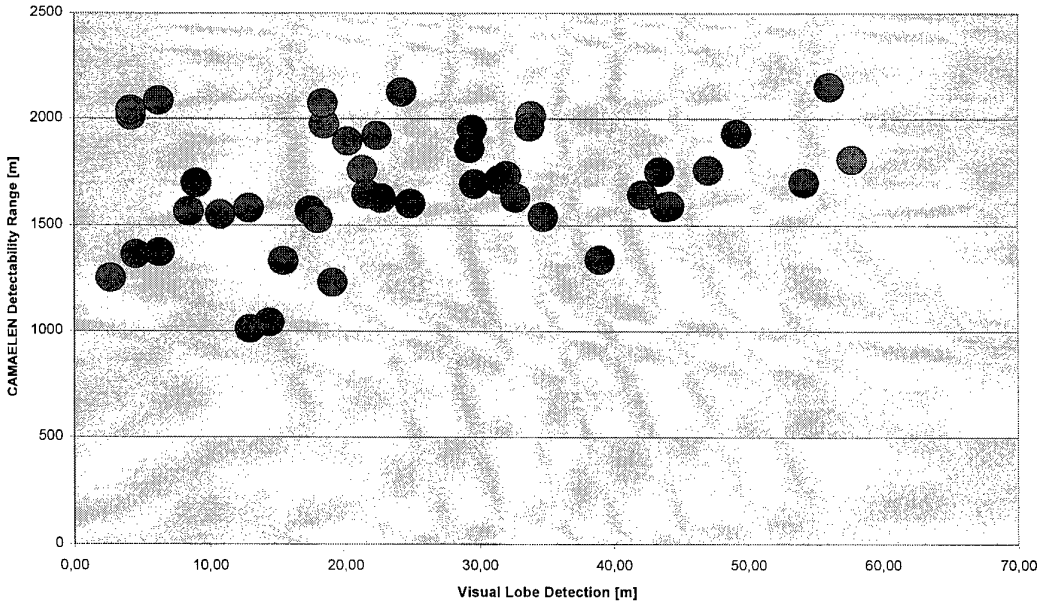


Fig. 9 CAMAELEON Detectability Range – Visual Lobe Identification („real word")

the targets on the screen. Because of the wide variety of distances the images were taken there also was a wide variety of object sizes on the screen, from which of course the visual lobe and the detectabiltiy ranges depend on.

Fig. 6 and Fig. 7 show the diagrams of CDR (CAMAELEON Detectability Range) versus VLD (Visual Lobe Detection) and VLI (Visual Lobe Identification) respectively. As expected because of the wide variety of object sizes the correlation is very high (PEARSON r^2 = 84% and 92% respectively). Proceeding from the fact, that the PEARSON correlation between the Square Root of Object Size (on the screen) and VLD and VLI respectively is r^2 = 73% and 86%, the "gain" resulting from CAMAELEON compared with Object Size only is 11% and 6% respectively.

To really judge about a detection or detectability model and/or compare it with others from our point of view it is absolutely necessary to hold object size constant as far as possible, that is in this case to take all the images from the same distance (as has been done in the CAMAELEON validation field trials). Otherwise you get interfering effects with different cues (size, atmosphere, resolution, contrast, texture) which cannot be resolved afterwards.

4.2. Extrapolation to "real world" situation

Another approach to the Search Data is to extrapolate the CAMAELEON calculations for the "real world" situation, that is here to calculate with the object sizes in the distance the images were taken from.

In this case the calculated CAMAELEON detectability ranges have to be compared with the Tangens of the Visual Lobe multiplied by the Distance. By this way we get the "real" visual lobe (in meters, not angle), and the object sizes vary in the natural ratios. On the other side we have these interfering effects in the images mentioned above resulting from taking the images from different distances, which should be avoided for CAMAELEON calculations.

Fig. 8 and 9 show the diagrams of CDR_r (CAMAELEON Detectability Range "real world") versus VLD_r (Visual Lobe Detection "real world") and VLI_r (Visual Lobe Identification "real word") respectively. The data are divided in two subgroups, *red* for short distances (sd), *green* for long distances (ld) the pictures were taken from), to support the hypothesis, that the CAMAELEON results should be better for short distances (see section 4). The correlation results are listed in Table 2.

	Pearson r^2	Pearson r	Spearman
CDR_r-VLD_r total	0.45	**0.67**	**0.68**
CDR_r-VLD_r sd	0.61	0.78	0.59
CDR_r-VLD_r ld	0.42	0.65	0.69
CDR_r-VLI_r total	0.06	**0.18**	**0.25**
CDR_r-VLI_r sd	0.44	0.66	0.52
CDR_r-VLI_r ld	0.02	0.14	0.20

Table 2 Correlation results (explanation see text)

It seems as if the hypothesis is supported (except Spearman for Visual Lobe Detection), but significance may be low because of the small sample size of images evaluated. The correlation with Visual Lobe Identification is very low, but in this case the difference between short and long distances is much higher than according to visual lobe detection. It supports the assumption, that CAMAELEON would give better results with standardized high resolution images of the targets taken from small distances.

5. CONCLUSIONS

Of course the results are not satisfactory. This is partly due to the special demands CAMAELEON makes on the quality of images and on the method of field trail evaluation (detectability as defined above instead of visual lobe).

On the other side the CAMAELEON model lacks further important features which influence the detectability of targets as color and luminance level (CAMAELEON has problems with gloss for example) and shape primitives. Another problem which makes detection modeling so complex and is not solved at all - neither in the CAMAELEON model nor in any other detection models – are ***the interfering effects of different cues.*** In the moment the CAMAELEON features are combined in a simple detection probability model assuming independence of the different cues with constant weighting factors. This may be wrong, but is hard to analyze from the scientific and modeling side as well as from the validating side.

Starting point for further development is the question of the CAMAELEON model, that is the assessment of camouflage. So it is not intended to expand CAMAELEON to a detection model, which is able to calculate search time and visual lobe in an entire scene as defined for example in the Search Data.

Detection really also depends on parameters of the overall scene, but these cannot be influenced by camouflage in a narrower sense.

So further investigations instead will be done on the features already used, luminance level, color and shape primitives and their interactions concerning detectability.

6. REFERENCES

1. Hecker, R. , "Camaeleon - Camouflage assessment by evaluation of local energy, spatial frequency and orientation", in: *Proceedings SPIE Conference on Characterization, Propagation, and Simulation of Sources and Backgrounds II*, SPIE Vol. 1687, pp. 342-349, Bellingham, WA: SPIE, 1992.
2. Toet, A., Bijl, P., Kooi, F.L., Valeton, J.M., "A high resolution image data set for testing search and detection models", TNO-Report TM-98-A020, TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.

Assessing Camouflage Methods Using Textural Features

[1]Sten Nyberg and [2]Klamer Schutte

[1] Defence Research Establishment
Division of Command and Control Warfare Technology
P.O. Box 1165, S-581 11 Linköping, Sweden
Phone: (+) 46 13 378000
Fax: (+) 46 13 378252
E-mail: stenyb@lin.foa.se

[2] TNO Physics and Electronics Laboratory
Electro-Optical Systems
P.O. Box 96864, 2509 JG The Hague, The Netherlands
Phone: (+) 31 70 3740469
Fax: (+) 31 70 3740654
E-mail: Schutte@fel.tno.nl

1. SUMMARY

Developments in the area of signature suppression make it progressively more difficult to recognize targets. In order to obtain a sufficient low degree of false alarms it is necessary to observe spatial and spectral properties. There is a genuine need to use spatial properties when analyzing the difference between a target area and a background area. This is more relevant today since modern signature suppression techniques have focused on the reduction of distinct features, like hot spots in the infrared band. The approach is to apply texture descriptors to characterize the background and also more or less camouflaged targets. In addition, other descriptors are used to characterize man made objects. It is necessary to focus on features which discriminate targets from the background, and this demands a more precise description of the background and the targets than usual. The underlying assumption is that an area with more or less observable targets has different statistical properties from other areas. Statistical properties together with detected target specific features like straight lines, edges, corners or perhaps reflections from a window have to be combined with methods used in data fusion. Experiments with a computer program that estimates the statistical differences between targets and background are described. These differences are computed using a number of different distance measures.

44 images from the Search_2 image data set [20] are used and mean search time and number of hits are predicted using textural features. The long term goal is to find methods for assessing signature suppression methods, especially in the infrared wavelength area.

Keywords: Terrain, texture, camouflage, assessment, optical, infrared, signature suppression

2. INTRODUCTION

This paper describes work done in an attempt to characterize the spatial variations in natural backgrounds. There is a genuine need to use spatial properties when analyzing the difference between a target area and a background area. This is more relevant today when modern signature suppression techniques are often used to reduce more distinctive features like hot spots in the infrared band which used to be sufficient. The approach here is to apply texture descriptors to characterize the background and also to the more or less camouflaged targets. In addition, other descriptors are used to characterize man made objects. These often have straight lines and edges.

Using texture information together with other kinds of information such as multispectral and temporal features makes the analysis and the assessment possible of signature reduction methods, reconnaissance systems, optical countermeasures, weapon sights and target seekers.

The literature contains attempts to performance assessment of signature suppression techniques [1]. However, there is still a need to find good methods. Many make assumptions that sometimes are difficult to verify.
In the future, the developments in the area of signature suppression will make it more and more difficult to recognize targets. In order to obtain a sufficient low degree of false alarms it is necessary to observe spatial and spectral properties. Also motion, if present, is an important feature. It is necessary to focus on features that discriminate targets from the background, and this demands a more detailed description of the background than usual. If time is not critical an approach using geometrical models is preferable. Given limited time and resolution one has to rely on measuring selected features. The underlying assumption is that an area with more or less observable targets differs in statistical properties from background areas. Statistical properties together with detected target specific features like straight edges, corners or perhaps reflections from a window have to be combined with methods used in data fusion. Experiments with a computer program estimating the statistical differences between targets and background are described. The long term goal is to find methods for assessing signature suppression methods, especially for infrared, but also for visual wavelengths.

Several ways to analyze images make it possible to assess different methods of signature reduction. One way is to visualize the properties of an image region in different ways.

- Displaying the Wiener spectrum (another name for power spectrum) for a region of interest. Specific features may show up in such an image.
- Displaying some relevant image transformations, like edge or line images.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

- Displaying a Wiener spectrum for a small region around every pixel in the image. In this case it is easier to examine local events in the image.

- Compute parameters that describe different features of the Wiener spectrum, like shape and distribution as examples of descriptors.

- Using one or several feature measures to define some kind of similarity measure or the opposite distance measures.

- Compute some measures that combine (uncamouflaged or camouflaged) target and background information.

Visualization of feature images is important because it is sometimes impossible to condense all the information down to a single number. Just like in image quality, color or texture analysis, several dimensions are needed to characterize a situation accurately. However, to validate these measures, there is a big demand for simple figures like detection time or signal-to-noise ratio.

An often-used method to visualize the similarity of a given set of features is trying to isolate targets from their background. In this case the image is segmented in target areas and background areas.

The ultimate validation is of course to test a method in real life in a target detection experiment. Using images of the scenes, the process can be simulated with a computer. Having a large enough set of images it is possible to assess probability of detection and also for example false alarm rates etc.

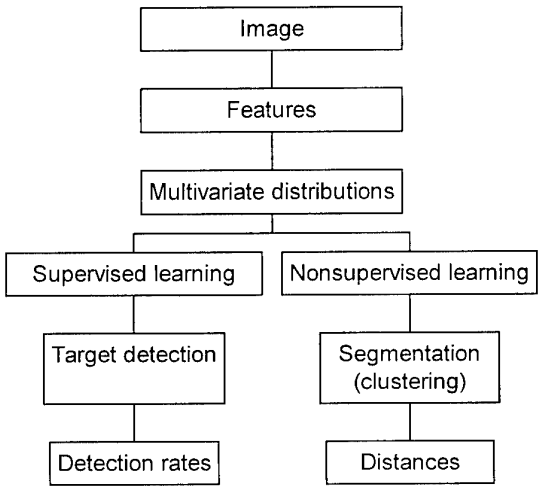


Figure 2.1 *Steps used in assessing differences between target and background.*

Figure 2.1 shows the different steps included when trying to find out which features are useful for the description of target and background properties. Several topics in figure 2.1 are discussed later.

Previous work in trying to find measures to assess camouflage effectiveness includes an investigation [2]. Some contributions in the literature are found [3-6], but none has yet come up with a sufficient method. A major problem is the lack of a good theory handling target detection in a cluttered environment. Theoretical work is often limited to the use of normal distributions for the background description. In a low observable situation this is completely unsatisfactory.

3. FEATURES

There are lots of texture measures in the literature. Designing a good set of features could be done using wavelet functions [7]. These are more or less limited in space and frequency domains. However, computing lots of wavelet functions is quite computationally expensive.

Tamura [4] has studied the relationship between textural features and visual perception. The six features he used were coarseness, contrast, directionality, linelikeness, regularity, and roughness. He found good correspondence in a ranking test with an implementation of 16 typical digitally computed texture measures. Woodroof [8] has estimated that three features should be enough to characterize normal textures. Texture measures based on the Fourier transform are shown in [5].

It is important to find features that are useful when trying to quantify the difference between targets and background.

Relevant properties for man made targets are given in the following section.

Table 3.1 Characteristic features for manmade targets.

Simple features:
- straight edges - homogeneous regions - specular reflection (from a planar surface) - homogeneous glints (from a uniform surface) - circular structures (=wheels)
Compound features:
- non-fractal (when one zooms in towards a part of a terrain scene, finer and finer details emerge. This will not happen to the same degree when looking at man made targets) - parallel edges - edge with at least one homogenous side - corners (= junctions of edges)
Motion properties:
- vibrations - tracks
Thermal features properties:
- "Hot spots" (from for example exhaust pipe) - tracks - heat from the motor engine, gun barrel etc
Spectral features:
- variations in reflected radiance and self-radiance

Table 3.2 Characteristic features for natural background.

Terrain background (texture features):
- coarseness
- contrast
- directionality
- linelikeness
- regularity
- roughness
Others:
- Fractal (when one zooms in towards a part of a terrain scene, finer and finer details emerge)
- non-stationary properties

Several features can be computed. Which features are useful to compute will be addressed later. About half of the features are based on image primitives like edges and blobs, that characterize targets. When computing these features the image is first for every pixel treated with an operator. The resulting image is then processed by a lowpass filter or something similar. This is done to find properties like concentrations of edges per region. For the background features, a local Wiener spectrum is computed for a region centered at each pixel. To save computation time, it is not necessary to compute the Wiener spectrum at every pixel. A coarse grid complemented with interpolation is adequate in most cases. In general, a good estimate is obtained if the grid separation is one fourth of the region size. When computing most of the features, a masking function may be applied to each local region to avoid boundary effects. It corresponds to an aperture function often used in spectral estimation. Here we use a very simple one, the Gaussian.

3.1 Target related features

The target related features used are: mean value, standard deviation, edge concentration, blob concentration, spoke maximum and edge coherence.

Although being first order statistics the mean value (*mean*) and standard deviation (*dev*) are included as they correspond to often used measures.

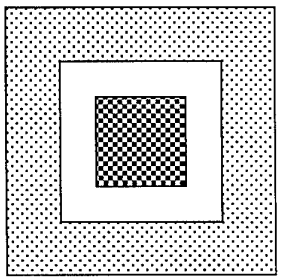


Figure 3.1 *The inner and outer mask used for computation of the blob concentration.*

The blob operator (*blob*) is defined with the help of figure 3.1. The mean values for the inner window and the outer window are computed and the difference is used as feature value if it exceeds a certain low threshold. Due to the sharp boundaries of these windows, the blob operator has to be applied to every pixel in the input image. As a texture measure for the local region, the mean value of the operator output is computed in the region.

The edge concentration (*edgeconc*) measure is the number of edge pixels in a local region around the center pixel. Edge-based texture measures have been investigated by Pietikäinen and Rosenfeld [9]

The spoke operator (*spokemax*), as described in [10], is shown in figure 3.2. It consists of eight spokes and is applied to every pixel in the image. Based on in how many spokes an edge segment is present, as represented in figure 3.2 by an arc, the presence of a small circular object may be detected. The output is an image where the pixel value corresponds to the number of hits that occur. Eight hits indicates a more or less closed curve, while three or four hits may indicate a corner. Instead of computing the mean value, the maximum value for each local region is computed.

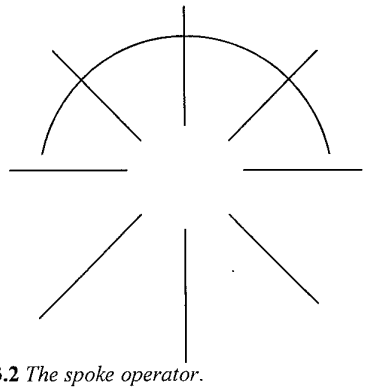


Figure 3.2 *The spoke operator.*

The implementation of the edge coherence (*edgecoh*) follows the method given in [11]. Other work in the same direction includes [12,13]. Its purpose is to indicate close parallel edges. Like the edge concentration feature, the edge image is used as input. Instead of summing the edge pixels for any direction, here only edges lying along the principal direction are summed. If the direction for an edge element differs from the principal direction it is weighted with respect to the difference in direction. If the edge magnitude is denoted *magn* then the edge coherence is computed according to,

$$edgecoh = (magnc - csumt)/csumn$$

where
magnc = edge image value in the center of the region

$$csumt = \sum(magn \cdot \cos(dirdiff))$$

$$csumn = \sum(magn)$$

and *dirdiff* = difference in direction between the center pixel and the others.

3.2 Background related features

The background features are all based on the Wiener spectrum, which is the squared magnitude of the local Fourier transform, and is called power spectrum in signal processing. They are isotropy, autocorrelation length, fractal dimension, directional autocorrelation, main direction, shape, low, medium and high frequency band energy, angular deviation, angular entropy and Fourier transform energy.

Given the spatial frequencies f_x, f_y and the Wiener spectrum magnitude $magn_{fx,fy}$ the isotropy is defined as in [14]

$$isotropy = 255 \cdot \frac{|sumu - sumv|}{\sqrt{(sumu + sumv)^2 - 4 \cdot sumuv^2}}$$

where

$$sumu = \sum \left(f_x^2 \cdot magn_{fx,fy} \right)$$

$$sumv = \sum \left(f_y^2 \cdot magn_{fx,fy} \right)$$

$$sumuv = \sum \left(f_x \cdot f_y \cdot magn_{fx,fy} \right)$$

When computing the autocorrelation length, basically the Wiener spectrum is integrated in the angular dimension. Only the frequency magnitude f_{xy} of the spatial frequency is used. The feature is defined as

$$autocorr = 10.0 \cdot f_{ny} \cdot \frac{msum}{fsum}$$

where

$$msum = \sum magn_{fx,fy}$$

$$fsum = \sum \left(f_{xy}^2 \cdot magn_{fx,fy} \right)$$

$$f_{xy} = \sqrt{f_x^2 + f_y^2}$$

f_{ny} = Nyquist frequency (=half the sampling rate)

Fractal geometry is a popular area for describing terrain and landscape. In addition, fractal dimension and lacunarity are two properties that can be computed [15]. Fractals for texture analysis have been studied by Gårding [16] and others. The Wiener spectrum is again treated as a function of the magnitude of the frequency. The fractal dimension is estimated from the Wiener spectrum magnitude using a least square fit of an angular integrated Wiener spectrum.

The lacunarity (*fracterr*) represents the amount of deviation an image exhibits from being fractal. Here it is a measure of how good a line will fit to the angular integrated Wiener spectrum.

The three features directional autocorrelation (*dirautoc*), mean direction (*eigenmean*) and shape (*shape*) are computed using a mass model of the Wiener spectrum and computing the inertia ellipsoid. The latter is computed by solving the eigenvalue problem

$$A \cdot I - \lambda \cdot A = 0$$

where A = covariance matrix with components a_{ij}. Here the Wiener spectrum is used as a distribution function.

Solving the eigenvalue equation gives two roots, λ_1 and λ_2 which correspond to the major and minor radius of the inertia ellipsoid.

The directional autocorrelation feature is defined as

$$dirautoc = const \quad \lambda_1$$

The main direction is defined as the direction of the principal axis of the inertia ellipsoid.

The shape feature corresponds to the elongeness of the inertia ellipsoid and is defined as the ratio between the minor and the major radius

$$shape = const \bullet (\lambda_1 - \lambda_2)(\lambda_1 + \lambda_2)$$

The next three texture measures, low, medium and high frequency band energy are probably the most relevant features when the problem is to characterize the scale of a pattern. The Wiener spectrum is summed in three different frequency bands. If the Nyquist frequency is f_{ny}, then the frequency limits are

lowband: 0 to $f_{ny}/4$

midband: $f_{ny}/4$ - $f_{ny}/2$

highband: $f_{ny}/2$ - f_{ny}

Figure 3.3 shows the summation areas.

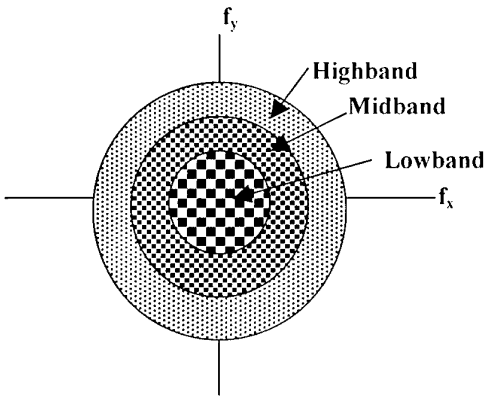


Figure 3.3 *Summation areas when computing Lowband, Midband and Highband.*

The total Fourier transform energy is simply defined as

$$ftenergy = k \cdot magnsum$$

where k is a constant and

$$magnsum = \sum \log(magn + 1) \log(dc)$$

magn = Wiener spectrum magnitude,

dc = magnitude at zero frequency.

A high value in ftenergy means that the image has a high degree of variation.
Knowing that the Wiener spectrum often falls of very rapidly with frequency, the use of logarithms gives high frequencies more weight.

3.3 Feature examples

Figure 3.4 shows an image divided into square grids of local regions and the corresponding Wiener spectra. Normally the regions are highly overlapped, with a center distance of one or two pixels.

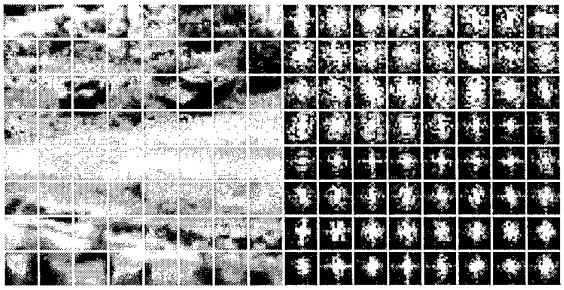


Figure 3.4 *Local spectra for a typical image. From left: the input image divided into the regions and the local spectra.*

The different background features relate to properties of these spectra. A few examples of feature images are given in figure 3.5.

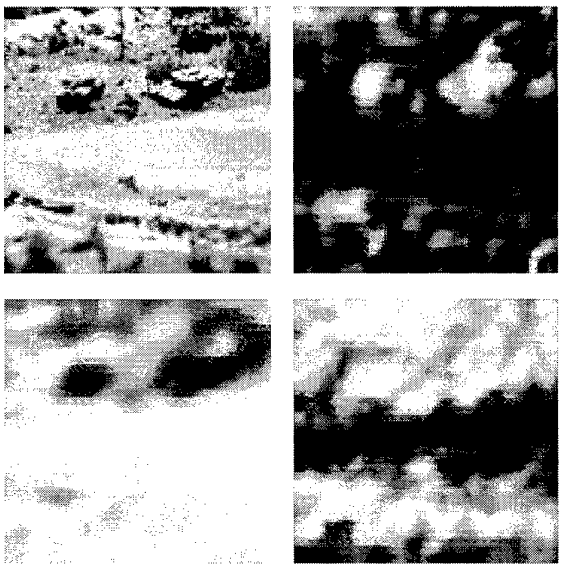


Figure 3.5 *An image (upper left), isotropy (upper right), autocorrelation length (lower left) and medium frequency band (lower right).*

4. DISTANCE MEASURES

We want to be able to express the difference between two areas as a distance using a space defined by some of the previously described features. The distance measures, see [2,21], have different underlying assumptions concerning the feature distribution. If mean values and standard deviations are used to characterize a feature, the distribution is normally assumed to be Gaussian and the features are assumed to be independent. Some distances used fall in this class. The reason for this is the simplifications made when applying them in practice. By using the covariance matrix, dependent features can be handled and the Mahalanobis distance is an example of this class. The Wilks measure uses no assumptions.

Because of a high degree of correlation between the different measures, it is advantageous to use distance measures that do not assume independent variables. Using this assumption leads to incorrect results.

Often it is of interest to use well-known quantities that have been used for a long time. One such measure is the signal-to-noise ratio (SNR) which is very common in connection with electrical signals. It is not easy to define an useful SNR for images, but attempts have been made by many researchers.

The different distance measures may be divided into three groups depending on how an area for target or background, is characterized. Most common is to use mean value and standard deviation.

Some measures take explicit consideration to dependent features. The Mahalanobis distance uses the covariance matrix to characterize one area and uses a feature point for the other. The original formulation of the Bhattacharrya measure makes no assumption about the target and background statistics, but often an approximation is used, where the distributions are assumed to be Gaussian and separable. The Wilks measure is a measure of similarity, which make no assumptions.

In table 4.1 the different measures used are listed.

Table 4.1 Listing of several distance measures.

#	Distance	Comments
1	Wilks	Parameter free
2	Bhattacharrya	May be parameter free
3	Mahalanobis	Uses the covariance matrix
4	Yaki	
5	Disabs	Only mean values
6	Dissqr	Only mean values
7	Tsnr	
8	dT_sum	
9	dT_suma	
10	dT_rss	
11	dT_rss4	
12	Doyle	
13	Doyle_mod	Includes a constant
14	Doyle_log	Includes a constant
15	Doyle_hybrid	Includes a constant

4.1 The distances

4.1.1 Wilks

The following description is given by Liu and Jernigan [14]. Let x_{igk} be the i:th feature value for the k:th sample of class g, where $i = 1,2,\ldots,m$ and m = the number of extracted features; $g=1,2,\ldots,G$ (G classes) and $k=1,2,\ldots,n_g$ (number of samples in class g). $N = \Sigma\, n_g$ is the total number of samples. The Wilks statistic is a measure of class separability that depends on within class and between class scatter matrices. The within class scatter matrix, W, and between class scatter matrix, B, are defined as

$$W = \left[w_{ij} \right]_{m \times m}$$

and

$$B = \left[b_{ij} \right]_{m \times m},$$

where

$$w_{ij} = \sum_{g=1}^{G} \sum_{k=1}^{n_g} \left(x_{igk} - \bar{x}_{ig} \right) \left(x_{jgk} - \bar{x}_{jg} \right)$$

and

$$b_{ij} = \sum_{g=1}^{G} n_g \left(\bar{x}_{ig} - \bar{x}_i \right) \left(\bar{x}_{jg} - \bar{x}_j \right).$$

$\bar{x}_{ig}$ and $\bar{x}_i$ are the mean value of class g and the total sample mean value for the i'th feature

$$\bar{x}_{ig} = \frac{1}{n_g} \sum_{k=1}^{n_g} x_{igk}$$

$$\bar{x}_i = \frac{1}{N} \sum_{g=1}^{G} \sum_{k=1}^{n_g} x_{igk}$$

The sum of within and between class scatter is the total scatter matrix T

$$T = W + B = \left[t_{ij} \right]_{m \times m}$$

$$t_{ij} = \sum_{g=1}^{G} \sum_{k=1}^{n_g} \left(x_{igk} - \bar{x}_i \right) \left(x_{jgk} - \bar{x}_j \right).$$

The Wilks statistics is the ratio of within class scatter to total scatter; $U = |W| / |T|$

4.1.2. Bhattacharrya

This is a measure of the overlap between two normalized distributions. If the distributions are $f(x)$ and $g(x)$, the Bhattacharrya coefficient b_{coeff} is defined as [17].

$$b_{coeff} = \int \sqrt{f(x) \cdot g(x)} dx$$

This quantity is related to false alarms and false detections.

In one implementation the features from the two regions to be compared are assumed to be Gaussian with mean values μ_1, μ_2 and standard deviations σ_1, σ_2. Assuming independent features gives the sum of the Bhattacharrya distance for the features between the two areas 1 and 2. Defining b as -log(b_{coeff}) gives

$$b = \frac{1}{N} \cdot \sum_{feat} \left(\frac{\left(\mu_{1,feat} - \mu_{2,feat} \right)^2}{4 \cdot \left(\sigma_{1,feat}^2 + \sigma_{2,feat}^2 \right)} + 0.5 \cdot \log \left(\frac{\sigma_{1,feat}^2 + \sigma_{2,feat}^2}{2 \cdot \left(\sigma_{1,feat} \cdot \sigma_{2,feat} \right)} \right) \right)$$

where the summation is done over all the features used.

4.1.3 Mahalanobis distance

This distance often occurs in connection with normal distributions. It is a measure from one point in a distribution to the center of the distribution. It is defined as [18].

$$r = \sqrt{(x - \mu)^t \Sigma^{-1} (x - \mu)}$$

Here x is the feature for a point in the image and the feature for rest of the image is characterized by the mean value μ and the covariance matrix Σ. Sometimes a small target area is compared with a larger background. In this case the target area statistics is approximated by its mean value and used for x in the expression above.

4.1.4. Yaki

This measure was designed by Yakimovski [19] in order to find out whether two regions are of the same kind or not. He found a measure, here called *yaki* for simplicity, that is for one feature given by

$$yaki = \left(\sigma_{12} \right)^2 \sigma_1 \cdot \sigma_2$$

where σ_{12} = standard deviation of the feature in the union of region 1 and region 2
σ_1= standard deviation of the feature in region 1
σ_2= standard deviation of the feature in region 2
Assuming Gaussian models for the two regions with mean values of μ_1 and μ_2 and standard deviations of σ_1 and σ_2 then the above expression may be evaluated to give

$$yaki_{feat} = 1 + \frac{(\mu_1 - \mu_2)^2}{(4 \cdot \sigma_1 \cdot \sigma_2)} + \frac{(\sigma_1 - \sigma_2)^2}{(2 \cdot \sigma_1 \cdot \sigma_2)}$$

Sometimes the constant 1 in the above expression is neglected in order to make the yaki measure look like a signal-to-noise ratio. If several independent features are used this measure will be given by

$$yaki = \sum yaki_{feat}$$

where $yaki_{feat}$ is computed for each feature according to equation above.

4.1.5. T-Student snr

In one application there was a need for simple measures that were fast to compute and has similarities to simple known measures, in this case the signal-to-noise ratio. The T-Student test [52] is used to see if two distributions are similar. We define it as

$$tsnr = |\mu_1 - \mu_2| \left(\sigma_1^2 + \sigma_2^2 \right)$$

Using mean values and standard deviations means that the underlying distributions are assumed to be normal.

4.1.6. Disabs

$$Disabs = \frac{1}{N} * \sum_{Features} \left(|\mu_T - \mu_B| \right)$$

4.1.7 Dissqr

$$Dissqr = \frac{1}{N} * \sum_{Features} \left((\mu_T - \mu_B)^2 \right)$$

4.1.8 dT_rss

$$dT\_rss = \frac{1}{\sqrt{N}} \bullet \sqrt{\sum_{Features}\left((\mu_T-\mu_B)^2+\sigma^2{}_T\right)}$$

4.1.9. dT_rss4

$$dT\_rss4 = \frac{1}{\sqrt{N}} \bullet \sqrt{\sum_{Features}\left((\mu_T-\mu_B)^2+4\bullet\sigma^2{}_T\right)}$$

4.1.10. dT_suma

$$dT\_suma = \frac{1}{N} * \sum_{Features}\left(|\mu_T-\mu_B|+|\sigma_T-\sigma_B|\right)$$

4.1.11. dT_sum

$$dT\_sum = \frac{1}{N} \bullet \sum_{Features}\left(|\mu_T-\mu_B|+\sigma_T\right)$$

4.1.12. doyle

$$doyle = \frac{1}{\sqrt{N}} \bullet \sqrt{\sum_{Features}\left((\mu_T-\mu_B)^2+(\sigma_T-\sigma_B)^2\right)}$$

4.1.13. doymod

$$doyle\_\mathrm{mod} = \frac{1}{\sqrt{N}} \bullet \sqrt{\sum_{Features}\left((\mu_T-\mu_B)^2+k\bullet(\sigma_T-\sigma_B)^2\right)}$$

where k=0.412.

4.1.14. doylog

$$doyle\_\log = \frac{1}{\sqrt{N}} \bullet \sqrt{\sum_{Features}\left((\ln(\mu_T)-\ln(\mu_B))^2+k\bullet(\ln(\sigma_T)-\ln(\sigma_B))^2\right)}$$

where k=0.00477.

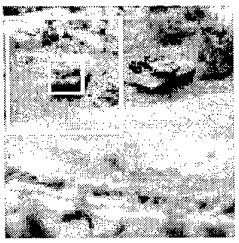
4.1.15. doyhyb

$$doyle\_hybrid = \frac{1}{\sqrt{N}} \bullet \sqrt{\sum_{Features}\left((\ln(\mu_T)-\ln(\mu_B))^2+k\bullet(\sigma_T-\sigma_B)^2\right)}$$

where k=0.000023.

4.4. Examples

An example of distance computation is shown in Figure 4.1. The distances are chosen in an earlier experiment.



DISTANCES

Mahalanobis 2.5
T_student_snr 1.0
Bhattacharrya 0.6
Yakimowski 0.8

Figure 4.1 *Distance computation using the isotropy and autocorrelation length. The inner area outlines the target area. The background area is defined as the area between the inner and outer square.*

Since many measures are used in the comparisons in a later part they will be defined here. The order here is in no way indicating their relevance.

Other examples of distance computations are given in section 5 and 6.

5. EXPERIMENTS WITH THE SEARCH_2 IMAGE DATA SET

44 images from the Search_2 data set [20] have been used in some experiments trying to correlate the distances from several distance measures with perceptual measures on detection time and hits performance. The images were limited in field-of-view to have a size of 256*256 pixels. They were selected with a magnification such that the target width occupied around 25 to 50 pixels. 5 images are from the B1 set, 26 from the B4 set and 13 from the B16 set. The tables in this section summarize experiments using several features and several distance measures. In several cases an exhaustive search has been performed to find the highest correlation with the perception data. Ideally, a model would be derived beforehand, to limit the search to relevant cases.

Table 5.1. Correlation between distance and detection time.

Rank	Features	Distance	Correlation
1 feature			
1	Isotropy	dTsum	0.653
2	Autocorr	dT_rss4	0.638
3	Isotropy	dT_rss	0.634
2 features			
1	Dirautoc, isotropy	dTsum	0.737
2	Dirautoc, isotropy	dT_rss	0.709
3	Ftenergy, isotropy Dirautoc, isotropy	dT_rss4 dTsuma	0.705
3 features			
1	Edgecoh, dirautoc, isotropy	dTsum	0.756
2	Edgecoh, ftenergy isotropy Dirautoc, isotropy, medfreq	dT_rss4 dT_rss	0.728
3	Dirautoc, isotropy, lowfreq	dTsuma	0.708

The three best are shown, just to indicate that there is no big difference between the good ones in each experiment. Using several features gives a better result but the risk is to adjust to the current image data set too much. In table 5.1 and 5.2 the distances are correlated with the detection times.
Some experimentation showed that correlation with the inverse of the distances gave a somewhat better result. The corresponding results are shown in table 5.3 and 5.4. A nonlinear function may be used, but again an adjustment to the current data set has to be avoided. The correlation is given with three decimals in the tables just to indicate small differences. In practise only the first decimal may be relevant.

Table 5.2. Correlation between inverse distance and detection time.

Rank	Features	Distance	Correlation
1 feature			
1	Isotropy	dTsuma	0.751
2	Isotropy	dTsum	0.740
3	Isotropy	dT_rss	0.730
2 features			
1	Dirautoc, isotropy	dTsuma	0.877
2	Dirautoc, isotropy	dTsum	0.856
3	Dirautoc, isotropy	doyle	0.852
3 features			
1	Dirautoc, isotropy, shape	dTsuma	0.880
2	Dirautoc, edgecoh, isotropy	dTsum	0.863
3	Dirautoc, edgecoh, isotropy Dirautoc, isotropy, mean	doyle dissqr	0.857

Table 5.3. Correlation between distance and hits.

Rank	Features	Distance	Correlation
1 feature			
1.	Ftenergy	dT_rss4	0.612
2	Ftenergy	dT_rss	0.571
3	Autocorr	dTsum	0.563
2 features			
1.	Ftenergy, isotropy	dT_rss4	0.688
2	Ftenergy, isotropy	dT_rss	0.649
3	Dirautoc, isotropy	dT_sum	0.639
3 features			
1	Dirautoc, ftenergy isotropy	dT_rss	0.695
2	Ftenergy, isotropy, shape Edgecoh, ftenergy, isotropy	dT_rss4	0.692
3	Autocorr, Dirautoc, isotropy	dTsum	0.688

Table 5.4. Correlation between inverse distance and hits.

Rank	Features	Distance	Correlation
1 feature			
1.	Autocorr	disabs, dissqr	0.687
2	Ftenergy	dT_rss	0.684
3	Shape	mahala	0.681
2 feature			
1.	Dirautoc, isotropy	dTsuma	0.783
2	Dirautoc, ftenergy	dT_rss	0.765
3	Autocorr, dirautoc	dT_sum	0.754
3 features			
1	Fracterr, highfreq, isotropy	mahala	0.803
2	Fractdim, ftenergy, isotropy Autocorr, dirautoc, mean	dTsuma dissqr	0.802
3	Highfreq, isotropy, mean	doylehyb	0.796

5.1. Comments

The tests indicate that the best result will be obtained using mean and variance based distances. Also it is evident that the inverse distance gives a better correlation reaching up to 0.85 in some case. The different tests also indicate that the features isotropy and dirautoc are among the best to use. If a third feature will be used, then ftenergy is a natural choice. One reason that isotropy is good is that it reacts to small straight edge segments that are common on targets but unusual in the background.

Better results may perhaps be obtained if the whole scene is processed. Now there is no estimation of possible false alarms outside the small background area used.

6. APPLICATION TO CAMOUFLAGE ASSESSMENT

Figure 6.1 shows a sequence of images where the targets are more and more camouflaged (simulated here by lowering the target contrast). The features used are directional autocorrelation distance (dirautoc) and isotropy. In the scatter image to the right of the image the covariance ellipses for the target area are plotted

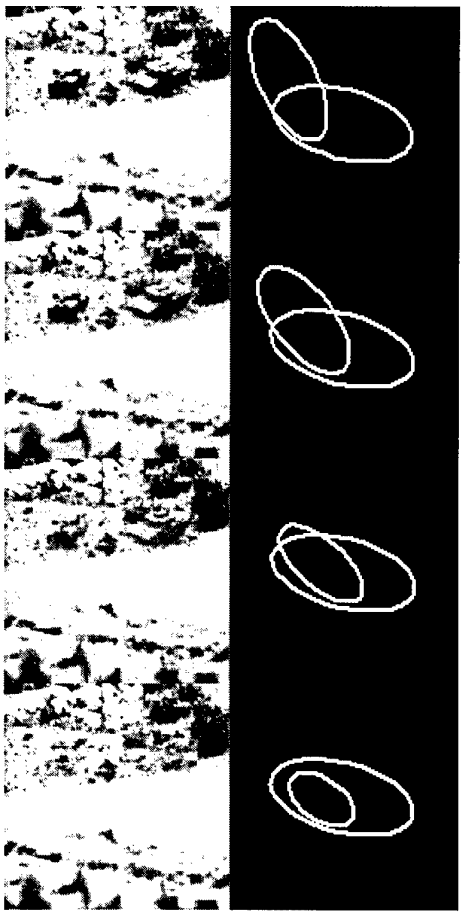


Figure 6.1 *A sequence of images where the targets are more and more camouflaged (simulated here by lowering the target contrast). To the right of the images scatter plots are shown with covariance ellipses for the target area and background area.*

It can be seen from the ellipses in the middle column that the overlap between the two ellipses increases as more and more camouflage is applied.

Several distance measures are computed for each camouflage level. Table 6.1 shows the distances. An earlier similarity measure was computed as the overlapping area from the two covariance matrices. This common area could be interpreted as the Bhattacharrya distance if the distributions for target and background are uniform and restricted to the covariance ellipse. However, the usual Bhattacharrya measure performed better.

Table 6.1 *Distances for different degrees of camouflage. Image 1 is the upper image in figure 6.1 and image 4 is the lower one.*

Measure	Distance			
	Image 1	Image 2	Image 3	Image 4
Mahalanobis	2.116	1.393	0.609	0.586
T_student_snr	1.143	0.779	0.434	0.395
Bhattacharrya	0.777	0.376	0.148	0.231
Yaki	0.882	0.403	0.161	0.272
DisAbs	54.151	36.760	17.749	12.496
DisSqr	54.524	37.192	19.185	16.069
Similarity				
Wilks	0.729	0.857	0.966	0.967

7. DISCUSSION

Using image analysis techniques it is possible to obtain a measure of similarity between camouflaged targets and the surrounding areas. It is also possible to compare targets having different degrees of camouflage with background areas. The difficult task is to select a suitable set of features to use.

Future work might also include integration of spatial properties with spectral and temporal features. This is necessary if assessment of a given signature suppression method is to be done. Furthermore the distance measures have to be "calibrated", for example related to recognition distances. A step in this direction is the experiments with the Search_2 image data set. The price to be paid by using many features is a heavy computation load, a disadvantage that will be less relevant in the future.

Experiments with the Search_2 image data set indicates the need for some model in order to find relevant features. However, it seems possible to use quite a few features to get reasonable result. In these experiments the features isotropy and directional autocorrelation distance seem to give some useful results. The tests indicate that the best result will be obtained using mean and variance based distances. Also it is evident that the inverse distance gives a better correlation reaching up to 0.85 in some case.

Better results may perhaps be obtained if the whole scene was processed. Now there is no estimation of possible false alarms outside the small background area used. This indicates that a multivariate normal distribution may be a possible model for the Wiener spectrum.

8. REFERENCES

1. Gerhardt, G., R., Meitzler, T., J. Performance assessment methodology for ground vehicle infrared and visual signature countermeasures (CMs), SPIE, Vol. 1687, pp 334-341, 1992.

2. Nyberg, S, Uppsäll, M, Bohman, L, An approach to assessment of camouflage methodology, SPIE proceeding no 1967: Aerospace Sensing, April 1993,

3. Reed, T. R., du Buf, J., M., H.. A review of recent texture segmentation and feature extraction techniques, CVGIP, Image Understanding, Vol. 57, No. 3, May, pp. 359-372, 1993.

4. Tamura, H., Mori, S., Yamawaki, T. Textural features corresponding to visual perception, IEEE Transaction on systems, man, and cybernetics, VOL. SMC-8, No. 6, pp. 460-473, June 1978.

5. Stromberg, W., D., Farr, T., G. A Fourier-based textural feature extraction procedure, IEEE Transactions on geoscience and remote sensing, VOL. GE-24, NO. 5. pp 722-731, September 1986.

6. Werman, M., Peleg, S., and Rosenfeld, A. A distance metric for multidimensional histograms, CVGIP, No 32, pp 328-336, 1985.

7. Bovik, A., C., Clark, M., Geisler, W., S. Multichannel texture analysis using localized spatial filters, IEEE Transactions on pattern analysis and machine intelligence, VOL. 12, No. 1, pp 55-73, January, 1990.

8. Woodruff, C., A proposed methodology for the spatial characterization of foliage backgrounds, Department of Defence, Technical note MRL-TN-510, AD-A178 747, 1987.

9. Pietikäinen, M., K., Rosenfeld, A. Edge-based texture measures, IEEE Transaction on systems, man, and cybernetics, VOL. SMC-12, No. 4, pp. 585-594, July/August 1982.

10. Minor, L., G., Sklansky, J., The detection and segmentation of blobs in infrared images, IEEE Transactions on Systems, Man, and Cybernetics, VOL. SMC-11, No. 3, pp. 194-201, March 1981.

11. Rao, A., R., Schunck, B., G. Computing oriented texture fields, CVGIP, Vol. 53, No 2, pp 157-185, 1991.

12. Bigün, J., Local symmetry features in image processing, PhD thesis, Diss. No. 179, Linköping University, Sweden 1988.

13. Kass, M., Witkin, A. Analyzing oriented patterns, CVGIP, Vol. 37, pp 362-385, 1987.

14. Liu, S., Jernigan, M., E. Texture analysis and discrimination in additive noise, CVGIP, No 49, pp 52-67, 1990.

15. Keller, J. M., Chen, S. Texture description and segmentation through fractal geometry, CVGIP, Vol. 45, pp 150-166, 1989.

16. Gårding, J. A note on the application of fractals in image analysis, Proceedings of SSAB (Swedish Society for Automated Image Analysis, Lund, Sweden, pages 80-83. 1988.

17. Duda, R., Hart, P., Pattern classification and scene analysis, John Wiley & Sons, p 40, New York, 1973.

18. Duda, R., Hart, P., Pattern classification and scene analysis, John Wiley & Sons, p 24, New York, 1973.

19. Schachter, B., J., A survey and evaluation of flir target detection/segmentation algorithms. pp 49-57, AD-A120 072, 1983.

20. Toet, A., Bijl, P., Kooi, F.L., Valeton, J.M., A high-resolution image data set for testing search and detection models, TNO-report TM-98-A020, April 1998.

21. Copeland, A., C., Trivedi, M., M., McManamey, J., R., Evaluation of image metrics for target discrimination using psychophysical experiments, Opt. Eng. 35(6), pp 1714-1722, June 1996.

Image Discrimination Models for Object Detection in Natural Backgrounds

A. J. Ahumada, Jr.
NASA Ames Research Center
Mail Stop 262-2
Moffett Field, CA 94035
U.S.A.
E-mail: aahumada@mail.arc.nasa.gov

1. SUMMARY

This paper reviews work accomplished and in progress at NASA Ames relating to visual target detection. The focus is on image discrimination models, starting with Watson's pioneering development of a simple spatial model and progressing through this model's descendents and extensions. The application of image discrimination models to target detection will be described and results reviewed for Rohaly's vehicle target data and the Search 2 data. The paper concludes with a description of work we have done to model the process by which observers learn target templates and methods for elucidating those templates.

Keywords: target detection, image discrimination models, video quality metrics, target template learning, response correlation images, Cortex Transform, Discrete Cosine Transform, Minkowski summation

2. INTRODUCTION

The vision research laboratory at NASA Ames Research Center has developed a number of image discrimination and video discrimination models. Although these models are not themselves models for the search and detection of targets in complex scenes, they can be used to estimate the visibility of a target in a fixed, unchanging background. For some applications, this task may be a useful simulation of the detection task. Also the visual representation components of the models may be taken out and incorporated in more complete models of the search and detection situation.

An image discrimination model takes as input a pair of images and gives as output a number relating to the probability that the observer will be able to discriminate the difference between the two images. In our models each image is converted to a visual representation and then the difference between the two visual representations is aggregated using a Minkowski summation index. Differences among our models mainly result from different visual system representations.

3. MODEL REVIEW

3.1. Watson' s Simple Spatial Model

The first true image discrimination model was developed by Watson [1]. The basic element of the model was the linear sensor element with a Gabor receptive field similar to that of a simple cell in primary visual cortex. These cells were assumed to occur in quadrature pairs and to be arranged in layers of units that were self-similar, but spaced 1 octave apart in spatial frequency. Because the sampling was approximately adequate to represent all the pictures in the image, when euclidean distance was used as the summation exponent and the model was space-invariant, its predictions were the same as any single linear filter model with the same contrast sensitivity function.

3.2. Watson' s Cortex Transform Model

The next major advance in image discrimination models was Watson' Cortex Transform model [2]. The basic elements of this model are still linear orientation selective filters, but they are computed by means of a filtering scheme cleverly designed to be a pyramid transform. The transform domain amplitude is quantized according to a nonlinear just-noticeable-difference scale to provide automatic image compression [3]. This nonlinear scale provides a masking mechanism. If the background image has raised the amplitude level of a transform coefficient before a signal is added, the signal amplitude must be larger to cause a just-noticeable increase in the coefficient amplitude. The image quality metrics of Daly [4] and Lubin [5] are close decendents of this model. Watson also developed a version of the model that is much more computationally efficient by using the Discrete Cosine Transform as a crude approximation of the Cortex Transform [6].

3.3. Masking from other "cortical units"

Foley [7] has shown that not all masking can be explained by the nonlinear response of simulated cortical units using psychophysical measurements of Gabor targets masked by gratings of different frequencies and orientations. Watson and Solomon [8] developed an image discrimination model where units neighboring in space, spatial frequency, and orientation contribute to a divisive inhibition of each other. This model returned to Gabor-shaped receptive fields for the original linear filters, taking advantage of the increased speed of new computers and ignores image reconstruction from the visual representation.

3.4. Simplified Models

The increased complexity of the models with between-unit masking lead us to try models that used a simple global RMS contrast to provide the masking [9]. Although this model is easily shown to be wrong in detail because it lacks selectivity in position and spatial frequency, it can provide surprisingly good approximations to standard masking results. A slightly more complex version of the model was contructed to allow for background images that are not constant in luminance and RMS contrast [10].

3.5. Image Sequence Discrimination Models

A sophisticated detection model will base its visual representation on the spatio-temporal retinal signal. Our labs have developed two discrimination models for video sequences. One is a temporal extension of the DCT-based image model [11]. The other is an extension of the simplified model for non-homogenous backgrounds [12]. Both use recursive filtering in the time domain. The second model [11] keeps separate representations for a "parvo" channel (high spatial resolution, low temporal resolution) and a "magno" channel (medium spatial resolution, high temporal resolution).

4. DISCRIMINATION MODEL APPLICATION

4.1. Previous Results

Rohaly, Ahumada, and Watson have compared several discrimination models or metrics on their ability to predict target detection performance in natural backgrounds [13]. The target detection task had several simplifications from realistic

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

detection tasks that made it suitable for the discrimination models. There was no search component to the task, the target if present was in the center of the image. For each target image a matched background image was made by replacing the target with a plausible section of background carefully blended with the original background. A discrimination experiment was run with the same monochrome, lower-resolution images that were given to the discrimination models. It showed that the detectability of the targets in the detection experiment, with different targets intermixed, was closely correlated to the discriminability of the reduced images with each target/no-target pair considered separately. Because of this, any visible difference between the images could be used as a basis for the detection.

Six models were tested. The first three were: 1) the difference between the images in the digital domain, 2) the difference between the images in the luminance contrast domain, contrast sensitivity filtered, 3) the Cortex Transform model (with a nonlinear amplitude transformation to provide within-unit masking). The next three were those model outputs normalized by a global contrast measure. For 4) the global contrast measure was simply the variance of the background image digital values. For 5) and 6) the contrast measure was the RMS contrast of the background luminance contrast image filtered with the contrast sensitivity function of the model.

The results were easy to remember. The order of the models, best to worse, with < meaning significantly better and = meaning approximately the same, was 4=5=6<3<1=2. Basically, the key to good model performance was masking. The two measures with no masking were the worst, the Cortex Transform model with within-unit masking was better, and any model with a global masking index was even better.

These results, and the results of some fixed noise masking studies with simulated airplane targets on runways, were the impetus for the development of the simplified model for nonhomogeneous backgrounds [10].

4.2. Search 2 Results

The single filter model with global masking was presented with small cut-outs of gray scale versions of the 44 Search 2 target images, together with matching versions with the targets removed [14].

The first step in the model calculations was to convert the images from digital gray scale, g, to luminance, Y, using the equation:

$$Y = 64.32((g-18)/(109.22+g-18))^{2.3} .$$

Next the images were converted to luminance contrast, C, using L0, the mean luminance of the image with the target removed.

$$C = Y/L0 - 1 .$$

The images were then filtered with a Difference-of-Gaussians contrast sensitivity filter. The center Gaussian had a spatial spread (1/e half width) of 2 arc minutes, the surround spread was 16 arc minutes, and the ratio of the surround volume to the center volume was 0.685. The filter was normalized to have a peak gain of unity in the frequency domain. The discriminability d' of the images is then estimated by the Euclidian distance between the filtered images, d, normalized by the Root-Mean-Square contrast of the filtered background image, c0.

$$d' = s\, d\, (1 + (c0/c2)^2)^{0.5} .$$

The parameter c2 is the masking threshold in contrast units and was set to 0.05. The sensitivity parameter s was set to give a contrast sensitivity of 114 for a filtered signal with constant unit contrast over an area of one square degree.

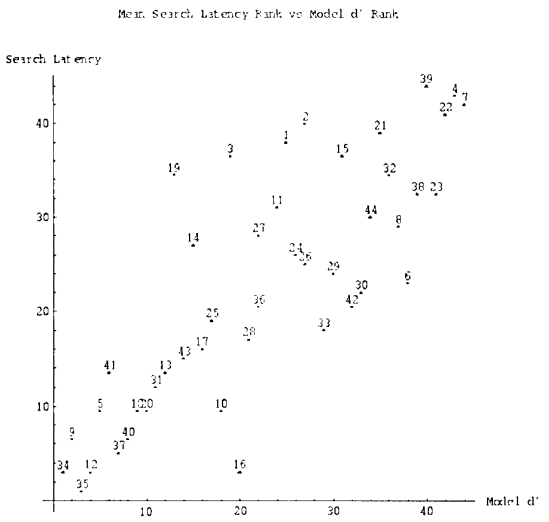


Figure 1

Figure 1 shows the ranks of the mean search latencies for the 44 signal images as a function the ranks of the model d's. The correlation coefficient for these ranks is 0.807, showing that this simple masking model predicts much of the variation in the search latencies.

This good performance is seen despite the fact that the model does not take into account (1) color differences, (2) target position, (3) object contours, or (4) texture differences.

The model does define and effectively combine (1) target size, (2) target contrast, and (3) background contrast variability.

It should be repeated that this is not a model of target detection. The hard part of the detection process was accomplished by the processes of limiting the image to the target region and replacing the target by a simulated background. For its simplicity, however, the model does a good job of capturing target visibility.

Two directions of improvement of this model are suggested. One is the addition of color. The other is a model that includes eccentricity. The model may have been improved by the poor color discrimination of the periphery. Also, in previous work with central targets we always found that a Minkowski distance with an exponent of 4 fit better than Euclidean distance, but for this data the Euclidean fit is better. This also may be a feature of peripheral search rather than foveal detection.

5. TARGET TEMPLATES

5.1. Fixed Noise vs. Random Noise Masking

Because we were interested in predicting detection performance in the presence of random noise, we have collected data comparing the relative effectiveness of random noise maskers and the fixed noise maskers that out image discrimination models are designed to predict [14]. A traditional signal detection approach to this problem predicts that random noise will be a stronger masker and that the difference should essentially depend on the ratio of the variability in the internal detection measure due to variability in the external noise to the internal variability in the measure when the noise is held constant. This internal variability was measured by Burgess and Colborne [17] for visual detection in noise by using the same noise on both intervals of a two-interval-forced-choice experiment. We call this method the twin noise method.

The interesting result comes from the comparison of twin noise with fixed noise. The standard models predict similar performance, but we find that fixed noise masking can be much less than twin noise masking [17].

5.2. Template Learning Models

To explain why fixed noise masking can be much less than twin noise masking we have developed models for template learning [18]. The basic idea is that in the fixed noise situation the observer is learning a template in a less variable situation, reducing the internal noise caused by variations in the template caused by the learning process itself. Another benefit of the fixed noise situation is that the template incorporates the fixed noise and reduces the spatial uncertainty in the detection process.

5.3. Template Identification Methods

A final research area that may be of interest to those trying to model target detection is our development of methods for identifying the features of the target that the observer is looking for. We add noise to the images in detection or discrimination tasks and correlate the noise pixels with the responses of the observers [19]. If the observer features are linear in the image pixel values, images of those features appear. If nonlinear features are used, the search is more tedious, but still possible [20].

6. ACKNOWLEDGEMENTS

I appreciate the assistance of Bettina L. Beard. This work was supported by NASA RTOP # 548-50-12.

7. REFERENCES

1. A. B. Watson (1983). Detection and recognition of simple spatial forms. In O. J. Braddick & A. C. Sleigh (Eds), *Physical and biological processing of images* (pp. 100-114). Berlin: Springer-Verlag.
2. A. B. Watson (1987a). The Cortex transform: rapid computation of simulated neural images. Computer Vision, Graphics, and Image Processing, **39**, 311-327.
3. A. B. Watson (1987b). Efficiency of an image code based on human vision. Journal of the Optical Society of America A, **4**, 2401-2417.
4. S. Daly (1993). The visible differences predictor: an algorithm for the assessment of image fidelity. In A. B. Watson (Ed.), *Digital Images and Human Vision* (pp. 179-206). Cambridge, Mass.: MIT Press.
5. J. Lubin (1993). The use of psychophysical data and models in the analysis of display system performance. In A. B. Watson (Ed.), *Digital Images and Human Vision* (pp. 163-178). Cambridge, Mass.: MIT Press.
6. A. B. Watson (1993). DCT quantization matrices visually optimized for individual images. B. Rogowitz and J. Allebach, eds., Human Vision, Visual Processing, and Digital Display IV, SPIE Proceedings, 1913, (SPIE: Bellingham, WA) pp. 202-216.
7. J. M. Foley (1994). Human luminance pattern-vision mechanisms: masking experiments require a new model. Journal of the Optical Society of America A, **11**, 1710-1719.
8. A. B. Watson & J. A. Solomon (1995). Contrast gain control model fits masking data. Investigative Ophthalmology and Visual Science, **36** (Suppl.), 438.
9. A. J. Ahumada, Jr. (1996). Simplified vision models for image quality assessment, J. Morreale, Society for Information Display International Symposium Digest of Technical Papers, Society for Information Display, Santa Ana, CA, 27, pp. 397-400.
10. A. J. Ahumada, Jr., B. L. Beard (1998). A simple vision model for inhomogeneous image quality assessment. J. Morreale, ed., Society for Information Display Digest of Technical Papers (Santa Ana, CA), 29, Paper 40.1.
11. A. B. Watson, J. Q. Hu, J. F. McGowan III,& J. B. Mulligan (1999). Design and performance of a digital video quality metric. In B. E. Rogowitz and T. N. Pappas, eds., Human Vision and Electronic Imaging IV, SPIE Proceedings, 3644, Paper 17.
12. A. J. Ahumada, Jr., B. L. Beard, R. Eriksson (1998). Spatio-temporal discrimination model predicts temporal masking functions. SPIE Proceedings, 3299, Paper 14, pp. 120-127.
13. A. M. Rohaly, A. J. Ahumada, Jr. & A. B. Watson (1997). Object detection in natural backgrounds predicted by discrimination performance and models. Vision Research, 37, pp. 3225-3235.
14. A. Toet, P. Bijl, F. L. Kooi, J. M. Valeton (1998). A high resolution image data set for testing search and detection models.TNO Report TM-98-A020, TNO Human Factors Research Institute, Soesterberg, The Netherlands.
15. A. J. Ahumada, Jr., B. L. Beard (1997). Image discrimination models predict detection in fixed but not random noise. Journal of the Optical Society of America A, 14, 2471-2476.
16. A. E. Burgess, B. Colborne (1988). Visual signal detection: IV. Observer inconsistency. Journal of the Optical Society of America A, 5, 617-628.
17. A. J. Ahumada, Jr., B. L. Beard (1997). Image discrimination models: detection in fixed and random noise. SPIE Proceedings, 3016, pp. 34-43.
18. B. L. Beard, A. J. Ahumada, Jr. (1999). Detection in fixed and random noise in foveal and parafoveal vision explained by template learning. Journal of the Optical Society of America A, 3, 755-763.
19. B. L. Beard, A. J. Ahumada, Jr. (1998). Technique to extract relevant image features for visual tasks. In B. E. Rogowitz and T. N. Pappas, eds., Human Vision and Electronic Imaging III, SPIE Proceedings, 3299, pp.79-85.
20. E. Barth, B. L. Beard, A. J. Ahumada, Jr. (1999). Nonlinear Features in Vernier Acuity. In B. E. Rogowitz and T. N. Pappas, eds., Human Vision and Electronic Imaging IV, SPIE Proceedings, 3644, Paper 8.

A CONTRAST METRIC FOR 3-D VEHICLES IN NATURAL LIGHTING

G. Witus
Turing Associates, Inc.
1392 Honey Run Drive
Ann Arbor, MI 48103
USA
E-mail: witusg@umich.edu

G. Gerhart
U.S. Army Tank-automotive and Armaments Command
AMSTA-TR-R / MS 263
Warren, MI 48397
USA
E-mail: gerhartg@tacom.army.mil

1. SUMMARY

Ground vehicles in natural lighting tend to have significant and systematic variation in luminance over the presented area. This arises, in large part, from the vehicle surfaces having different orientations and shadowing relative to the source of illumination and the position of the observer. These systematic differences create the appearance of a structured 3-D object. 3-D appearance is an important factor in search, figure-ground segregation and object recognition.

This paper presents a contrast metric based on the 3-D structure of the vehicle, and an analysis of search performance for the Search_2 imagery. The analysis employs the traditional P-infinity-times-negative-exponential model of search time distribution. P-infinity and mean search time are modeled as functions of the target signature. The signature metric is one over the product of vehicle size and contrast. The value of the metric is measured by the ability to account for variance in observed search performance.

The 3-D structure contrast metric performs better than RSS contrast, and both perform dramatically better than the area-weighted average contrast. Target height performs better than either target area or square root of area. The signature metric accounts for over 80% of the variance in probability of detection and 75% of the variance in search time as measured in the TNO perception tests. When false alarm effects are discounted, the metric accounts for 89% of the variance in probability of detection and 95% of the variance in search time. The predictive power of the signature metric when it is calibrated to half the data and evaluated against the other half, is 90% of the explanatory power.

Keywords: Contrast ratio, 3-D perception, computational vision model, shape from shading, target acquisition, search

2. INTRODUCTION

Size and contrast have long been used to characterize the signature of simple targets in simple scenes for the purpose of analyzing search time and probability of detection. Size and contrast have been found to be good predictors of search and detection performance for stylized 2-dimensional targets, such as uniform disks and 4-bar patterns, against uniform backgrounds [Blackwell, 1943] [Ratches, et al., 1975].

Unfortunately, the standard area-weighted average contrast ratio has not proven to be a good predictor of search and target acquisition performance for complex targets in complex scenes. D'Agostino, et al., [1997] suggested a variety of possible modifications to the area-contrast metric to account for statistical luminance variation within the target and local surround. Peli [1996] concluded that the common measures of contrast are inadequate to explain detection performance for Gabor patches against uniform backgrounds, and suggested a computational contrast metric based on multi-scale band-pass filtering as an alternative.

Ground vehicles in natural lighting present non-uniform appearance when the surfaces of the vehicle are at different orientations with respect to the source of illumination and the observer (see fig. 1). The differences in shading between the adjacent surfaces reveal the 3-D structure. The appearance of common vehicles, from typical perspectives, under natural lighting is readily learned. This contributes to the perception of a 3-D object at a location, recognition of characteristic structural features, and classification as a potential vehicle.

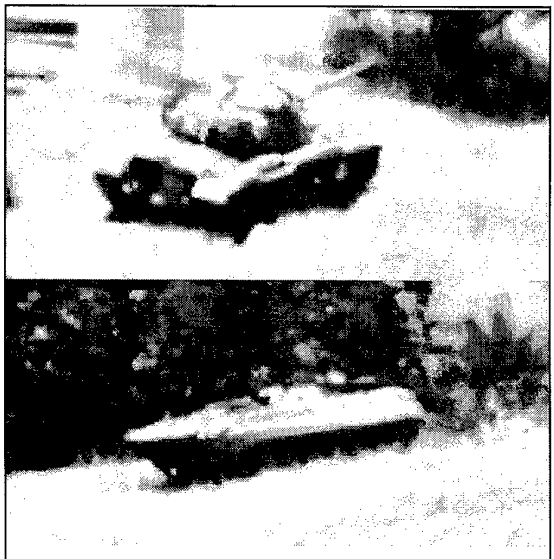


Fig. 1. Example images of vehicles showing distinctive shading on different surfaces.

This paper presents the initial results of exploratory research to develop a contrast metric based on the 3-D vehicle structure, in natural lighting, relative to typical observer perspective. The objective of this research was to determine if a contrast metric could be defined based on the vehicle 3-D structure that would produce improved predictions of

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

probability of detection and time to detect for military targets in natural backgrounds.

There is a substantial body of prior research suggesting that the perception of 3-D structure as a result of shape-from-shading is a significant factor in visual search and target acquisition. (Depth perception from visual parallax is insignificant at tactical ranges. For a stationary observer and a stationary target, shading and prior knowledge of vehicle appearance are the primary factors in 3-D shape perception.)

Marr [1982] coined the term "the 2½-D sketch" to refer to the perception of a 3-D structure from surface primitives. Sun and Perona [1996] showed that 3-D shading produced "pop-out" detection (i.e., response time independent of the number of distracters, indicative of pre-attentive parallel processing). They also showed that search became serial (time linear in the number of distracters) when the 3-D shading was removed even though the edge structure was retained. Tarr and Kersten [1998] concluded that the human visual system uses illumination angles to extract 3-D shape, and that illumination effects (including shadows) are modeled with respect to object shape, rather than simply encoded in terms of their effects in the image. Jonides and Gleitman [1972] and Mack and Rock [1998] both demonstrated that pre-attentive object recognition directs visual attention. Liu, Knill and Kersten [1995], and Liu and Kersten [1998] found that human efficiency exceeded 100 percent of an ideal 2-D observer for 3-D object classification. Moore and Cavanagh [1998] demonstrated that perception of 3-D shape is possible from limited surface shading information, given familiarity with the 3-D object. Ullman [1996] has shown that observers use 3-D operations to recognize familiar objects presented in novel orientations. 3-D surface matching is also an approach being pursued for automatic object recognition systems designed to work in clutter with partially occluded targets (e.g., [Johnson and Hebert, 1998]).

3. MODELING APPROACH

3.1 Contrast, Size and Signature Metrics

This exploratory investigation employed a simplistic, low-resolution approach. If 3-D shading is a significant factor in search and target discrimination, then the effects should be apparent even though coarse analytic techniques were used. If coarse analysis does not reveal an effect, then the effect, if present, is probably not strong enough to be worth addressing in search and target acquisition models.

The conceptual 3-D vehicle model was based on the 3 cardinal surface orientations of a rectangular solid (vertical front, vertical side, and horizontal top). While military vehicles are not rectangular solids, the 3-region geometric model can be adapted with a little work. The projected view of a vehicle was divided into the following three regions (see figure 2):

1. Front/rear. The near-vertical, negatively sloped or self-shadowed portion of the front (or rear depending on the presented aspect). For armored vehicles this includes the lower glacis, front track/tire, and turret-chassis gap. For trucks, this includes the front grill, front of the cab, and front of the tires.
2. Side. The near vertical (e.g., within 10 degrees), negatively sloped or self-shadowed portion of the side, including the sides of the tracks or tires.
3. Top. All horizontal and near-horizontal surfaces up to a slope of 80 degrees. This includes all the small miscellaneous objects and protrusions on the vehicle. It includes the the upper glacis, top deck, roof, rear deck, turret armor. It also includes the sloped rear roof of the HMMWV.

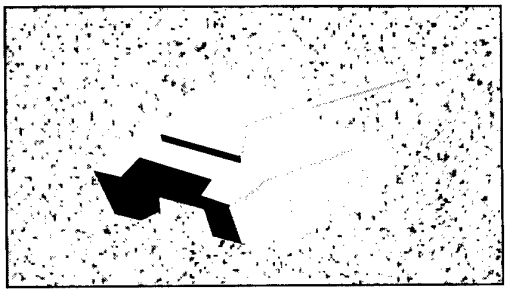


Fig. 2. Illustration of canonical front (rear), side, and top vehicle surfaces.

These canonical surfaces are meant to identify the major vehicle surfaces that typically have distinctive luminance resulting from different self-shadowing and angles relative to the observer and illumination. They address only the grossest level of 3-D structure. This level of resolution may be too coarse for modeling higher levels of target discrimination.

These regions also correspond to key structural features reported in field tests: darkly shadowed lower glacis, side profile, glint off roof or deck, lower grill, cab front, turret-chassis shadow. It is possible that the three surface orientations are significant because they correspond to important features for vehicle discrimination. It is also possible that the features are important because they reveal the 3-D structure.

The 44 Search_2 digital images [Toet, et al., 1998] were used in the demonstration analysis. All 44 images were used with no exceptions. The images were analyzed using Adobe Photoshop© to outline regions and compute gray-scale values. The local surround was taken to be a band surrounding the target with width equal to the target height.

The average gray-scale values for each region j, G_j, were converted to luminance values, L_j, via the calibration equation provided by Toet, et al., [1998]:

$$L_j = f(G_j) = 64.32 \left[(G_j - 18) / (G_j + 91.22) \right]^{2.3}$$

Since the calibration is a non-linear equation, a more accurate approach would have been to first convert pixel gray-scale to luminance, then compute the statistics.

The contrast for region j, C_j, is defined as the absolute value of the difference between the mean luminance of the region, L_j, and the mean luminance of the surround, L_{bkg}:

$$C_j = | L_j - L_{bkg} |$$

The vehicle contrast ratio metric, C_{veh}, is the area-weighted average of the contrasts of each of the three regions, divided by the luminance of the local background:

$$C_{veh} = \Sigma \, w_j \, Cj \, / \, L_{bkg}$$

where the weights, w_j, are the proportion of the presented vehicle area contributed by each region.

Two alternative contrast metrics were examined to provide a basis for relative comparison. These were the traditional area-weighted-average contrast [Ratches, et al., 1975] and the RSS contrast [D'Agostino, et al., 1997]. Both were computed by applying the non-linear gray-scale to luminance transform, f(), to statistics computed on the gray-scale images.

Signature metrics based on the area-weighted average contrast were uncorrelated with search performance (r^2 on the order of

0.3). Area weighted average contrast is not addressed in the remainder of this paper. This contrast metric was rejected.

The RSS contrast metric has been found to be an effective metric in other studies [D'Agostino, et al., 1997]. It is used as a reference for comparison with the 3-D structure contrast.

The RSS contrast ratio is the root-sum square of the target-background luminance difference and the target luminance standard deviation, divided by the mean background luminance:

$$RSS = [(\mu_{tgt} - \mu_{bkg})^2 + \sigma_{tgt}^{\ 2}]^{1/2} / \mu_{bkg}$$

For this comparison, the luminance standard deviation was estimated from the gray-scale mean and variance:

$$\sigma_{tgt} = [f([\mu_g^{\ 2} + \sigma_g^{\ 2}]^{1/2})^2 - f(\mu_g)^2]^{1/2}$$

where f() is the gray-scale-to-luminance calibration equation.

The signature metric, S_{veh}, used in the analysis is simply one divided by the product of the vehicle size measure, V_{veh}, and vehicle contrast measure, C_{veh}:

$$S_{veh} = 1 / (V_{veh} C_{veh})$$

The size metric in this analysis was the target minimum dimension, generally the vertical extent or height. Vehicle height was the measure of size used in the early Night Vision Laboratory target acquisition modeling [Ratches, et al., 1976]. Target height (vertical extent) was reported in the Search_2 documentation.

Two alternative size metrics have been proposed as alternatives to target minimum dimension: the vehicle presented area, and the square root of the presented area [D'Agostino, et al., 1997]. These size metrics were examined, but their performance was inferior to target height. Only analysis results using height are presented.

3.2 Search Model

The analysis employed the traditional search performance model that expresses probability distribution of detection over time as the product of a limiting probability of detection (P_{inf}) and a negative exponential distribution:

$$P_d(t) = P_{inf} * (1 - e^{-(t-\varepsilon)/Td})$$

where e is the minimum reaction time (nominally 0.5 seconds) and T_d is the mean time to detect given that a detection occurs [Washburn, 1981] [Ratches et al., 1975].

In the perception test subjects were given 60 seconds in which to search and respond [Toet, et al., 1998]. Toet reports the mean search time plus reaction time. To obtain the parameters of the search model, the effects of the 60-second response window and reaction time must be discounted. Assuming the negative exponential distribution of search time, given that a detection occurs, mean search time, discounting windowing and reaction time, can be computed from the reported mean search time, T_s:

$$T_d = (T_s - \varepsilon) / (1 - e^{-(60-\varepsilon)/Ts})$$

Toet et al. [1998] also reports the number of detections, N_d, false alarms, N_f, and misses, N_m. Probability of detection within 60 seconds can be calculated from this data:

$$P_d(60) = N_d / (N_d + N_f + N_m)$$

In the test image set, only one image had $P_d(60)$ less than 0.4, three images had $P_d(60)$ less than 0.5, five images had $P_d(60)$ less than 0.6, and 24 images had $P_d(60)$ greater than 0.95. Figure 3 shows the relative number of detection, false alarm and time-out (miss) responses in the perception test.

P_{inf} is computed from T_d and $P_d(60)$

$$P_{inf} = P_d(60) / (1 - e^{-(60-\varepsilon)/Td})$$

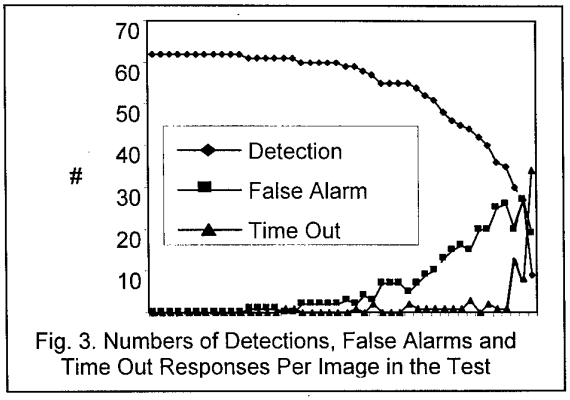


Fig. 3. Numbers of Detections, False Alarms and Time Out Responses Per Image in the Test

Given an unlimited search time, there are three possible outcomes: the observer can detect the target, false alarm, or conclude that there is no detectable target in the scene. Each of these is an absorbing state. As soon as the observer enters any one of these states, the search is over. Whenever a detection occurs, it is conditioned on having occurred before a false alarm and before the observer concludes that there is no target in the scene. In order for the conditional time to detect to have a negative exponential distribution, two criteria must be met:

(1) target detection, false alarms, and conclusion that no target is in the scene must be independent processes; and

(2) each of these processes must have a negative exponential distribution (albeit with different rates).

Under these conditions, the mean time to detect, conditioned on detection occurring first, is one over the sum of the individual rates of detection, R_d, false alarm, R_f, and concluding no detectable target is present, R_c:

$$T_d = 1 / (R_d + R_f + R_c)$$

P_{inf} is simply the ratio of the rate of true detection to the combined rates:

$$P_{inf} = R_d / (R_d + R_f + R_c)$$

These rates can be computed from the available data:

$$R_d = (1 / T_d) P_{inf}$$

$$R_f = (1 / T_d) N_f / (N_f + N_d)$$

$$R_c = (1 / T_d) - R_d - R_f$$

Cinlar [1975] and Washburn [1981] provide details of the mathematics of competing Markov processes.

The perception test in which the Search_2 data was collected used 35 mm slides with targets present in every scene. The subjects knew that each scene contained a vehicle. The subjects also knew that they had only 60 seconds in which to search the scene. Under these conditions, the subjects would presumably continue searching for the full 60 seconds. Since they knew a target was present, they would not conclude no detectable target was present within the first 60 seconds. This implies that R_c should be zero.

This hypothesis is supported by the data. The mean value of R_c, computed over the 44 images, is 0.0008, the maximum is 0.008, and the standard deviation is 0.0021. The expected time to conclude no detectable target is present ($1/ R_c$) is 21 minutes, and the standard deviation of the rate is 2.6 times the mean. R_c is not statistically significantly different from zero, and even if it was, it is so small as to be insignificant for this analysis. The remainder of the paper disregards R_c.

Figures 4 and 5 show the distribution of the rate of target detection, Rt, and the rate of false alarm, Rf. Note that the scales on the two graphs are an order of magnitude apart. Rt is greater than Rf for 43 of the 44 scenes.

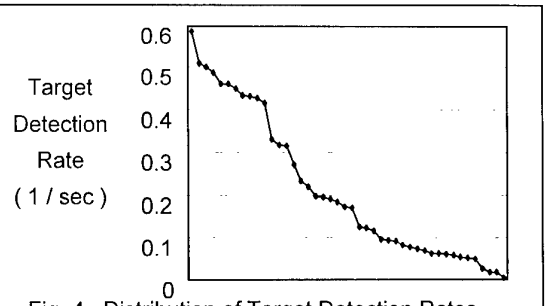


Fig. 4. Distribution of Target Detection Rates

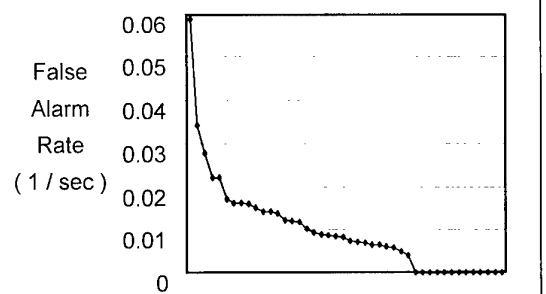


Fig. 5. Distribution of False Alarm Detection Rates

The mean time to detect a target, given that a detection occurs, discounting the effect of false alarms is

$$T_t = 1/R_d$$

When the effects of false alarms are discounted, P_{inf} has value one. The probability that a detection occurs within 60 seconds, discounting the effect of false alarms can be computed directly from the response data on the numbers of detections and missed targets

$$P_d(60 \mid \text{no false alarms}) = N_d / (N_d + N_m)$$

When the effects of false alarms are discounted, P_{inf} has value one. The probability that a detection occurs within 60 seconds, discounting the effect of false alarms can also be estimated from the computed from the unconstrained mean time to detect a target without false alarms, Tt:

$$P_d(60 \mid \text{no false alarms}) = (1 - e^{-(60-\varepsilon)Rd})$$

The root-mean-square (RMS) difference between these two estimates is 0.036, comparable to the sampling error in $P_d(60)$.

P_{inf} and T_d are modeled as simple linear functions of the signature metric. The model parameters (slope and intercept) are estimated from the data via linear regression. The related measures of search performance ($P_d(60)$, $P_d(60 \mid$ no false alarms), T_s and T_t) are modeled as functions of P_{inf} and T_d using the preceding search model equations.

4. ANALYSIS RESULTS

4.1 Gray-Scale Variance in the Vehicle Image

Partitioning the projection of the vehicle into the front, side and top regions accounted for 63 percent of the gray-scale variance over the entire target region. The area-weighted sum of the gray-scale variance within the three regions was 37 percent of the gray-scale variance over the entire target region. This indicates that these vehicle regions account for a significant proportion of the gray-scale variance in images of ground vehicles. Sources of residual variance include small features defining local surfaces with different orientations and self-shadowing, paint patterns, shadows from trees falling on the vehicle, and patches of foreground obscuration.

The RSS contrast metric includes all variance over the vehicle, regardless of structural significance or spatial scale of the variation. The RSS contrast and 3-D structure contrast have a strong statistical linear relationship ($r^2 = 0.90$).

4.2 Sampling Error in the Search Performance Data

Sampling errors are inherent to any test procedure with a finite number of subjects. If the identical experiment were repeated with different subjects, the results would differ due to sampling error and the stochastic nature of search and detection.

$P_d(60)$ is estimated as the proportion of observers correctly detecting the vehicle. Assuming observer responses are independent, the sampling error has a binomial distribution. For a given image, the one-sigma sampling error in $P_d(60)$ is given by the following equation:

$$\sigma_{Pd} = [P_d(60) * (1 - P_d(60)) / N]^{1/2}$$

where N is the number of subjects (N = 62).

Figure 6 shows a plot of sampling error in $P_d(60)$ versus observed $P_d(60)$ for the 44 Search_2 images. The RMS sampling error in $P_d(60)$ over the entire Search_2 image set is 0.0363. The standard deviation in measured $P_d(60)$ over the entire image set is equal to 0.187. Sampling error explains 3.8 percent of the variance in $P_d(60)$ over the image set.

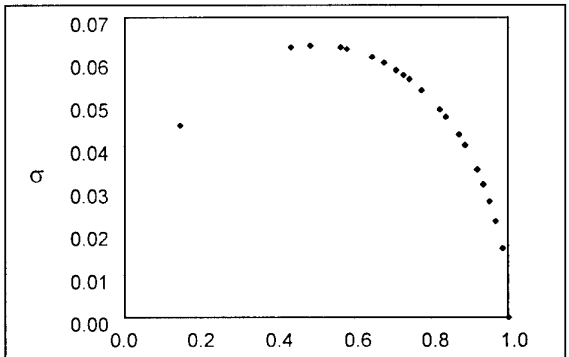


Fig. 6. $P_d(60)$ vs. Sampling Error in P_d (60).

The long-run probability of detection, P_{inf}, was not measured directly, but was computed from measured data. This makes the effects of sampling error difficult to compute. However the effects of sampling error can be approximated by assuming the random variables were measured. The sampling error in P_{inf} is 0.036, explaining 4.4% of variance.

The probability of detection in 60 seconds absent the effects of false alarms, $P_d(60 \mid$ no false alarms) is computed directly from the recorded data. The sampling error in $P_d(60 \mid$ no false alarms) is 0.025, explaining 3.8% of variance.

The mean search time reported, T_s, is a constant reaction time ($\varepsilon = 0.5$ sec) plus a random variable with a negative exponential distribution truncated at 60-ε seconds. For this analysis the standard deviation is approximated by the standard deviation of a negative exponential random variable with the same mean (i.e., no truncation). The standard deviation of a negative exponential random variable is equal to the mean. Since T_s is computed using response time data only for subjects who detect the vehicle, for any given image

the sampling error is equal to T_s divided by the square root of the number of subjects who correctly detected the vehicle:

$$\sigma_{Td} = (T_s - \alpha) / [\, N\, P_d(60)\,]^{1/2}$$

Figure 7 shows a plot of sampling error in T_s versus T_s for the 44 Search_2 images. The RMS sampling error in T_s over the entire Search_2 image set is 2.4 seconds. The standard deviation in measured T_s over the entire image set is equal to 7.58 sec. Sampling error explains 10.3 percent of the variance in T_s over the image set.

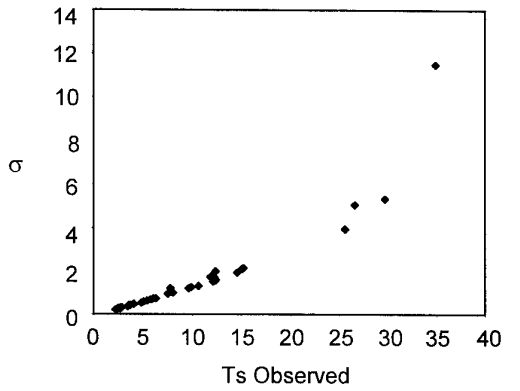


Fig. 7. T_s Obs vs. sampling error in T_s

The unconstrained mean time to detect, T_d, and the unconstrained mean time to detect in the absence of false alarms, T_t, were not measured directly, but were computed from measured variables. This makes the error due to sampling difficult to compute. The effects of sampling error can be approximated by assuming the random variables were measured. Both random variables have negative exponential distributions, so the 1-sigma sampling error is equal to the mean divided by the square root of the number of responding subjects. The estimated sampling error in T_d is 2.7 seconds, explaining 9.8% of variance. The estimated sampling error in T_t is 11.3 seconds, explaining 11.5% of variance.

4.3 Model Explanatory Power

The model has four free parameters that must be estimated from data: the slope and intercept of P_{inf} as a function of the signature metric, and the slope and intercept of T_d as a function of the signature metric.

The explanatory power of the model is measured by the percentage of variance in the observed search performance accounted for by the model. This is computed from the root-mean-square error between the model and observed data, and the variance in the observed data:

$$\%Var = 100\,(1 - RMS\_Error^2 / Observed\_Variance)$$

For a linear fit with parameters estimated via linear regression, the percentage of variance explained is equal to 100 times the Pearson correlation coefficient squared (r^2). Since the search model equations are non-linear, the percentage of variance accounted for is computed from the RMS error.

Figure 8 shows a scatterplot of the mean time to detect a target, given that the target is detected before a false alarm, but unconstrained by the 60 second time window of the experiment. The experimental value of T_d is computed from the measured search time. The model estimate of T_d is a linear function of the signature metric fit to the observed T_d.

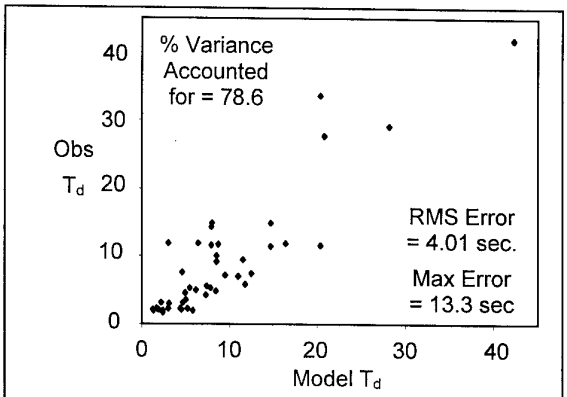


Fig. 8. Unconstrained Mean Time to Detect, T_d

Figure 9 shows a scatterplot of P_{inf} computed from the observed test data versus the linear function of the signature metric fit to the observed P_{inf} and truncated at one. Experimental values of Pinf are computed from T_d and raw response tallies.

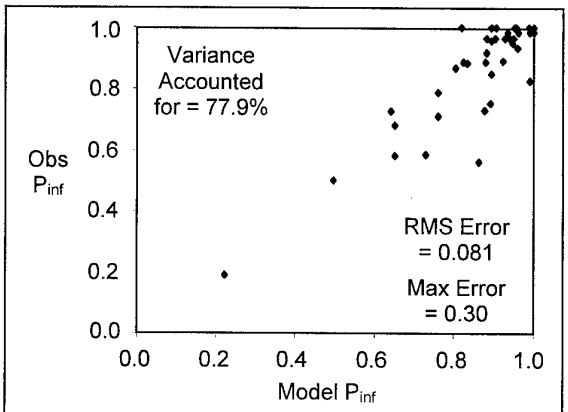


Fig. 9. P_d(infinity) P_{inf}

Figure 10 shows a scatterplot of the probability of target detection in 60 seconds computed directly from the tallies of observer detections, false alarms and misses, versus the model $P_d(60)$ computed from P_{inf} and T_d. The results are very similar to the Pinf results because, in most cases, the mean search time was much less than the 60 second response window.

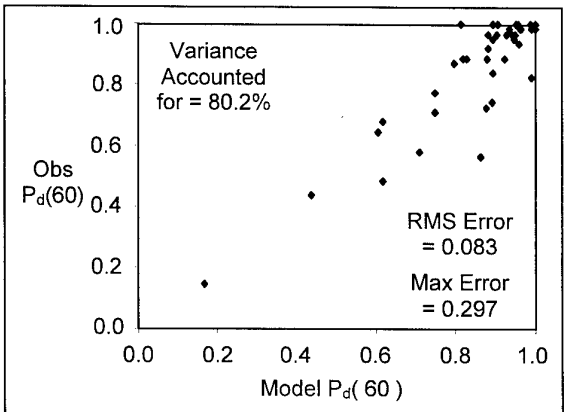


Fig. 10. P_d(60)

Figure 11 shows a scatterplot of the mean search time measured in the experiment, versus the mean search time calculated by the model accounting for the effects of competing false alarms and the 60 second response window. These results resemble the results for unconstrained search time because, in most cases, the mean search time was much less than the 60 second response window.

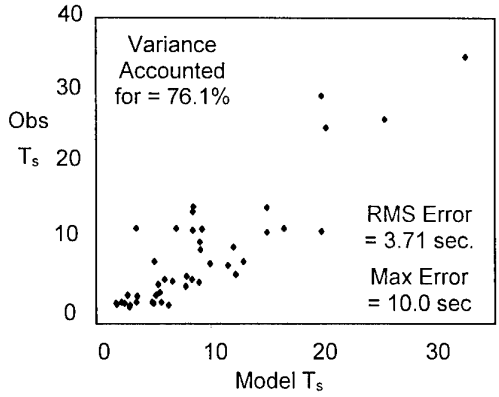


Fig. 11. Mean Search Time, T_s

Figure 12 shows a scatterplot of the mean time to detect a target without the requirement that the target detection occurs before a the first false alarm, T_t. It is the inverse of the rate of target detection. It is computed as T_d divided by P_{inf}. The experimental and model values of T_t are computed from the experimental and model values of values of T_d and P_{inf}. When the RSS contrast is used instead of the 3-D structure contrast, the percent of variance accounted for drops from 95% to 89%.

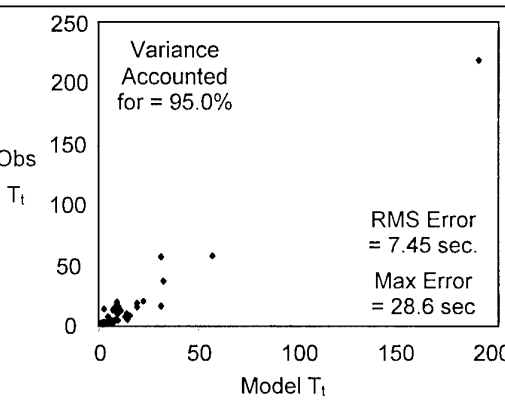


Fig. 12. Mean Detection Time Absent False Alarms, T_t

Many of the data points in figure 12 are clustered near the origin. The correspondence for low response time cases is seen more clearly when the logarithm of T_t is plotted (see figure 13). The logarithm operation is a non-linear transformation, so the percent of variance accounted for is different.

Interestingly, the percent of variance accounted for by Ln(Model T_t) is equal to the percent of variance accounted for by linear regression of the signature metric directly on Ln(Observed T_t). When the RSS contrast is used instead of the 3-D structure contrast, the percent of variance accounted for drops from 76% to 50%.

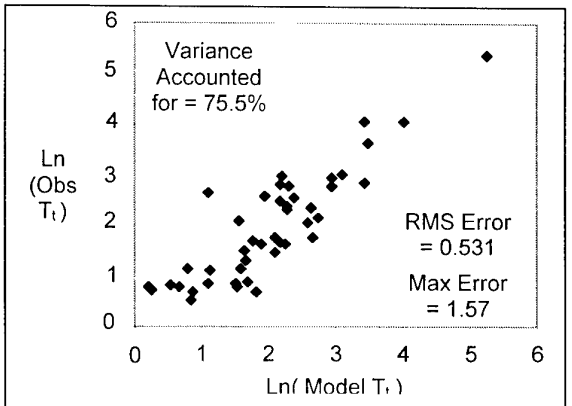


Fig. 13. Ln (Mean Detection Time Absent False Alarms), Ln(T_t)

Figure 14 shows a scatterplot of the probability of target detection in 60 seconds, without competition from false targets, i.e., excluding false alarms. The experimental value is computed from the tallies of detections and misses. The model value is computed from T_t.

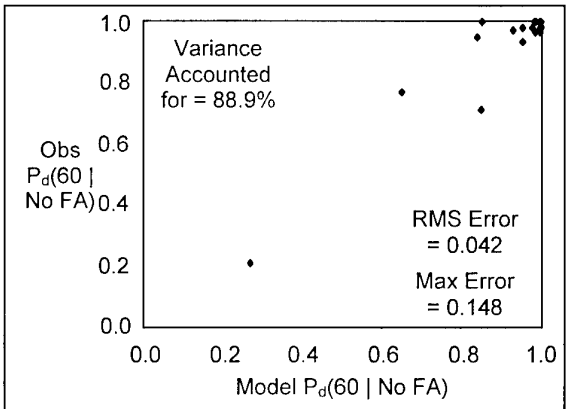


Fig. 14. P_d(60 | No False Alarms)

Tables 1 and 2 summarize the results of the comparison of the model to the data, and compare results obtained using the 3-D structure contrast metric with those obtained using the RSS contrast metric. Table 1 presents the percent of variance explained by the model. Table 2 shows the magnitude of the maximum model error.

Search Performance Measure	% Var 3-D	% Var RSS
Unconstrained Time to Detect, T_d	78.6	77.8
P_{inf}	77.9	76.0
Search Time, Ts	76.1	75.2
P_d(60)	80.2	78.6
Detection Time Sans F.A., T_t	95.0	88.5
P_d(60 | No False Alarms)	88.9	86.5

Table 1. Model Explanatory Power

Search Performance Measure	Max Error 3-D	RSS
Unconstrained Time to Detect, T_d	13.3	11.9
P_{inf}	0.30	0.31
Search Time, Ts	10.0	9.0
P_d(60)	0.30	0.31
Detection Time Sans F.A., T_t	28.6	60.2
P_d(60 | No False Alarms)	0.15	0.19

Table 2. Maximum Model Error

The results have several significant implications:

1. Signature metrics based on both the 3-D structure contrast metric and on the RSS contrast metric account for large proportions of the variance in search performance for this data set.
2. The 3-D structure contrast metric accounts for one to two percentage points more variance than the RSS contrast metric, except for the mean time to detect, absent false alarms where there is a 6.5 percentage points difference. This indicates that the 3-D structure contrast is a better measure observer response to the target signature. When the effects of false alarms are included, the additional variance due to this non-target source obscures the difference between the two contrast metrics.
3. The percentage of variance predicted by both metrics is significantly higher when the effects of false alarms are discounted. This is not surprising since the signature metrics do not measure potential false targets. The difference is greater for 3-D structure contrast than for the RSS contrast, further supporting the claim that 3-D structure contrast is a better measure of the effects of the target signature.
4. The difference in the percent of variance predicted with and without the effects of false alarms indicates the magnitude of the contribution of false targets to search performance variance. By this measure, false targets account for over 15% of the variance in the mean time to detect a target, and almost 9% of the variance in the probability of target detection within 60 seconds.
5. The maximum error in P_d(60 | no false alarms) is significantly lower that the maximum probability error when the effects of false alarms are not excluded.
6. The magnitude of the maximum detection time sans false alarms is large. However this error occurs at the one hard-to-detect image, for which T_t was 218 seconds. The error, as a percentage of the time for that data point, is 13% for the 3-D structure contrast metric and 28% for the RSS metric.

Several excursions were conducted to assess alternative vehicle size metrics. When the signature metric was calculated using the square root of the presented area instead of the target height, the percent of variance predicted was approximately 12 percentage points lower for P_{inf} and 3 percentage points lower for T_d. When the presented area was used, the results were 15 percentage points lower for P_{inf}, and 6 percentage points lower for T_d.

4.4 Signature Metric Measurement Error

Measurement error occurs because the original images were blurred. The boundaries of the vehicles and regions in the vehicles were not sharply delineated. This affected both the measurement of target vertical extend and luminance. Not only was the location of the boundary uncertain, but pixels near the boundary contained a mix of target and background luminance, or a mix of the luminance between two regions.

Two separate estimates of the 3-D structure contrast ratio were made. Toet et al., [1998] provided one measurement of target height. A second independent measurement was made in this study. These measurements provided two pair of independent measures of the signature metric. Each independent pair of estimates produced one estimate of the measurement error in the signature metric.

The one-sigma measurement error in the signature metric over the Search_2 image set is 0.016. Since the model is linear, the measurement error in the predictions of P_{inf} and T_d are 0.016 times the magnitude of the slope (-2.313 and 92.11 respectively). This analysis yields a one-sigma measurement error in the predictions of 0.036 for P_{inf} and 1.473 for T_d respectively. The measurement errors in the predictions of P_{inf} and T_d are less than the sampling errors in the perception test estimates of P_{inf} and T_d (0.036 and 2.7 seconds respectively).

In combination, the variance due to sampling error and signature metric measurement error together are 9.1 percent of the P_{inf} variance predicted by the model, and 18.5 percent of the variance in T_d predicted by the model. The predictive power of the signature metric cannot be the result of spurious sampling and measurement errors.

The signature metric is one over the product of the vehicle 3-D structure contrast and the vehicle height. Two measurements of height and contrast were made, to obtain two pairs of independent measurements of the signature metric. The two correlations between the two pair of signature metric measurements were 0.986 and 0.979. The sample standard deviation for a pair of independent measurements is simply the difference between them. The error estimate for two pairs of independent measurements is the RMS of the two estimates of the sample standard deviation.

Figure 15 presents a plot of the signature metric measurement error sample standard deviation versus the signature metric value. The correlation is 0.91, suggest a strong linear relationship with slope equal to 0.15. As expected, the measurement error is larger for small, low-contrast vehicles than for large, high-contrast vehicles.

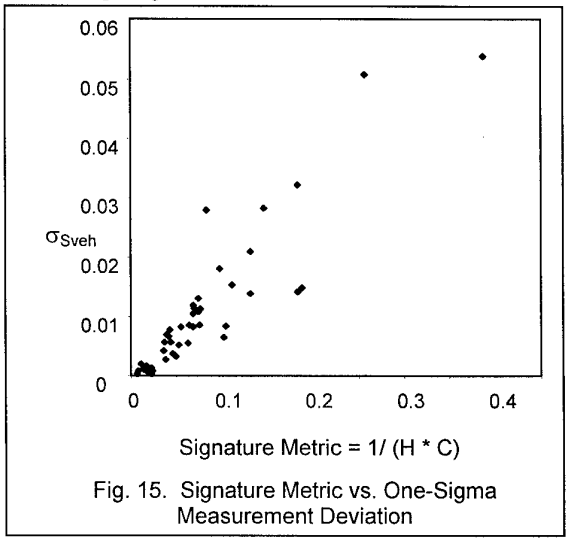


Fig. 15. Signature Metric vs. One-Sigma Measurement Deviation

4.5 Accounting for Residual Variance

The model accounts for 75 to 80 percent of the variance in the experimental data when the effects of false alarms are included, and 90 to 95% of the variance when they are discounted. This suggests that for the TNO Search_2 data and test, false alarms account for 10 to 15 percent of the variance in probability of detection and search time, respectively.

Sampling error accounts for approximately 4 percent of the variance in probability of detection and 10 percent of the variance in search time. Together the target signature, false alarms, and sampling error are sufficient to account for all of the variance in the experimental data. (It is not possible simply to sum the percentage of variance explained by sampling error with the percentage of variance explained by the signature metric because of spurious correlation when the model parameters were estimated from the data).

The signature metric was calculated by applying the non-linear gray-scale-to-luminance transform to the mean and RMS gray-scale values. The correct method is to apply the gray-to-luminance transform to the image, then compute statistics. This approximation may account for some of the residual variance.

The contrast metric did not include any measure of color contrast or texture differences. The metric did not address chromatic, luminance or contrast adaptation, or spatial filtering. The metric did not address the effect of the position of the target in the scene, or its position relative to other features that might attract of inhibit attention to the target location. These factors may contribute to the model error, but the effect is likely to be small because the unexplained error is small.

There is no term that can be added to the signature metric to account for the residual variance. The prediction errors in P_{inf} and T_d are only weakly correlated ($r^2 = 0.25$). This suggests that once target signature effects are discounted, probability of detection and search time are sensitive to different processes and/or have non-linear relationships with image characteristics.

4.6 Individual Effects of Size and Contrast

One over the 3-D structure contrast metric was modestly correlated with perception test data ($r^2 = 0.7$ for T_d and 0.6 for P_{inf}). The percentages of variance explained for T_t and P_d(60 | no false alarms) were 51% and 54%. The RSS contrast metric had comparable correlation to T_d and P_{inf}, but accounted for 10 percentage points less of the variance in T_t and and P_d(60 | no false alarms). The area weighted average contrast ratio had essentially no correlation with T_d or P_{inf}.

Target height, area and square root of area were only weakly correlated with T_d and P_{inf} (r^2 approximately equal to 0.4). Height had some correlation with T_t and P_d(60 | no false alarms) with r^2 on the order of 0.2. For area and square root of area, accounted for less than 10 percent of variance in T_t and and P_d(60 | no false alarms).

Height was less correlated with the 3-D structure contrast metric than it was with the RSS contrast metric ($r^2 = 0.3$ for the 3-D structure contrast metric, versus 0.4 for the RSS contrast metric).

These data indicate that height and 3-D structure contrast were largely independent dimensions, which individually were moderately correlated with search performance. Not surprisingly, their product was well-correlated with search performance. The same statements are true to a lesser extent for the RSS contrast metric.

4.7 Spurious Correlation and Predictive Power

When the same data are used to calibrate and evaluate the model, the percentage of variance accounted for is an accepted measure of explanatory (descriptive) power, but it is not truly a measure of the model's predictive power. In order to assess the model's predictive power, the model must be calibrated to one data set, then the prediction error evaluated for a separate, sequestered data set. This minimizes the effects of spurious correlation.

The Bootstrap statistical technique [Davison, 1997] was used to evaluate the predictive power of the signature metric. The Bootstrap technique involves repeated random partitioning of the data into two disjoint sets: the calibration data set containing, and the validation data set. The model parameters are estimated from the calibration set, then the RMS prediction error is calculated from the sequestered validation data set. Each partition produces an estimate of the variance predicted by the model.

In this particular application of the Bootstrap technique, the calibration and validation data sets each contained half of the data points. Twenty-two calibration data points were used in the linear regression to estimate the model parameters, and 22 data points in the validation set were used to measure the RMS error and the percent of variance in the validation data set predicted by the model. Two-hundred-fifty-two (252) random partitions were generated to compute the Bootstrap statistics.

The Bootstrap analysis was applied to investigate the ability of the signature metric to predict the logarithm of the mean time to detect a target in the absence of false alarms, T_t. T_t was chosen as the dependent variable because it had the clearest causal relationship to the signature metric. The logarithm of Tt was used because of the uneven distributions of observed T_t and for the signature metric (see figures 12 and 13).

Figure 15 shows the distribution of slope and intercept from the 252 partitions. The slope had median value 14.6, equal to the slope when all 44 points are used in the regression (theBootstrap mean and variance are 15.0 and 1.6 respectively). The intercept has a median value of 0.89 compared to 0.90 when all 44 points are used in the regression (the Bootstrap mean and variance are 0.88 and 0.14 respectively).

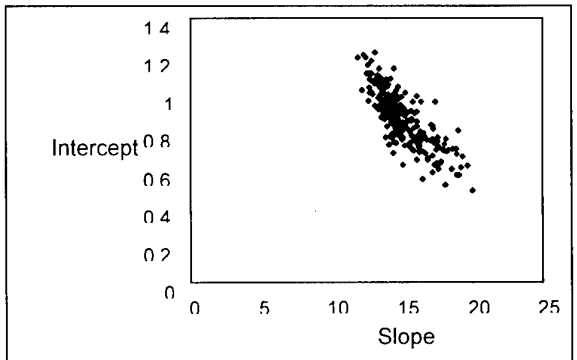


Fig. 15. Distribution of Calibration Parameter Values from Bootstrap Replications

Figure 16 shows the distribution of the percentage of variance in the calibration data sets versus the percentage of variance predicted in the validation data sets. The median percent of variance predicted in the validation data sets is 72% (the mean and variance are 70% and 10 percentage points, respectively). The median percent of variance explained in the calibration data sets is 78%, compared to 76% when all 44 points are used

in the regression (the Bootstrap mean and variance are 76% and 9 percentage points, respectively). On average (median and mean) the proportion of variance predicted in the validation data set is 92% of the proportion of variance accounted for in the calibration data set. This difference is due to spurious correlation, and indicates the difference between the explanatory and predictive power of the signature metric.

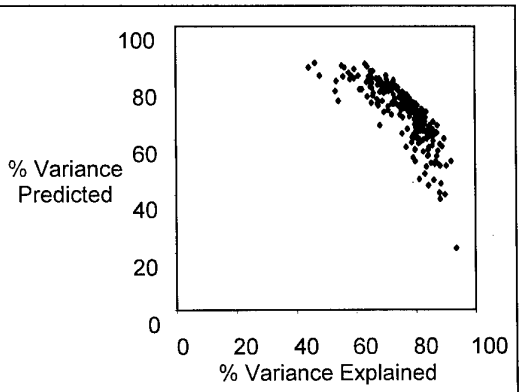


Fig. 16. Distribution of Percentage of Variance Explained and Predicted by the Signature metric

5. FINDINGS AND OBSERVATIONS

The traditional model of the distribution of search time was a useful model to analyze the experimental data.

With appropriate choice of definition of size and contrast, the simple signature metric equal to one over the product of size and contrast is a good fit to the observed data. It explains 75 to 80 percent of the variance in the test data, and 90 to 95 percent when the effects of false alarms are discounted.

The organization of the vehicles into three regions based on their orientation relative to the illumination and observer accounts for a significant portion of the gray-scale variance. Not surprisingly, the 3-D structure contrast metric and the RSS contrast metric are highly correlated and produce comparable results.

Nonetheless, the 3-D structure contrast metric is consistently superior to the RSS contrast metric, especially when the effects of false alarms are discounted. Variance due to false alarms obscures the difference in performance for the two contrast metrics. Both contrast metrics are far superior to the area weighted average contrast (which is no good at all).

Vehicle height is a better measure of target size, for use in product with a contrast metric, than either vehicle area or square root of vehicle area.

Responses to false targets, i.e., false alarms, account for a 10 to 15 percent of the search performance variance. Modeling the rate of false alarms as a function of the image properties has potential to improve search modeling.

There are a number of low-level and high-level visual phenomena not represented in this simple signature metric. Low-level factors include color contrast, chromatic and luminance adaptation, spatial filtering and contrast adaptation. Mid-level image processes include pre-whitening, edge detection and texture segregation. Beyond the vehicle structure, high-level (top-down) image properties include the location of the vehicles relative to terrain features that might attract attention or direct attention away, and position of the vehicle in the image.

These factors could account for the unexplained variance. However, they were not major contributors to search performance variance in the Search_2 image set. These factors could be more significant in other image sets containing greater variation on these dimensions.

The Search_2 vehicles do not present significant perceptible camouflage. Camouflage adds variance to the image. The RSS contrast metric will yield higher values for vehicles with camouflage than for vehicles without camouflage (assuming the same mean luminance), and thus will predict higher P_{inf} and lower T_d for camouflaged vehicles than for comparable non-camouflaged vehicles. The 3-D structure contrast metric is camouflage-neutral since it is based only on the mean luminance of different target regions and does not incorporate any higher-order statistics.

The Search_2 vehicles do, in some cases, have perceptible structures within the front, side and top regions. This increases P_{inf} and decreased T_d. These structures add variance to the image, which increases the value of the RSS contrast metric, which leads to higher predicted P_{inf} and lower predicted T_d. The 3-D structure contrast metric is neutral with respect to structures within the three regions.

Neither the 3-D structure contrast metric nor the RSS contrast metric are able to distinguish modulation due to internal structure from modulation due to camouflage or foreground obscuration (e.g., brush or nets). More sophisticated signature analysis is needed to make this distinction.

6. CONCLUSIONS

The 3-D structure of the vehicle is a promising basis for signature analysis. Basic research suggests that shape from shading and 3-D appearance are pop-out cues, focus visual attention, facilitate figure-ground segregation. This analysis provides evidence that 3-D structure is an important factor in search and target acquisition in natural settings.

The 3-D structure contrast metric was useful in analyzing the Search_2 image set. The simple approach explored in this paper may not be robust enough for a wide variety of image sets. Future research should explore extending the 3-D structure analysis approach and using it in combination with a computational model of front-end visual processing.

The traditional model of the distribution of time to detect a target was a useful framework with which to analyze search performance. The search model was extended to express P-infinity as a function of the rates of detection, false alarm and deciding that no detectable target is present. This extension enable the analysis to quantify the effects of false alarms on variance in search performance.

The signature metric had very strong explanatory power for this data set, especially when the effects of responses to false targets were discounted. Limited Bootstrap analysis suggests that the predictive power of the model is 92 percent of the explanatory power.

False alarms were a significant factor contributing to variance in search performance. Further research is needed to demonstrate effective models to predict the rate of false alarm from image properties and top-down knowledge.

The specific quantitative results of this analysis, especially the calibration of P_{inf} and T_d as linear functions of the signature metric, are unlikely to transfer to other perception tests and image sets. Observer response depends on the test-specific factors such as the proportion of images with no target, the relative penalty of false alarms versus missed detections, the response time window, search area, etc. Image sets that

contain different distributions of target signatures, false targets, scene complexity (terrain features), etc. will lead to different quantitative results.

7. ACKNOWLEDGEMENTS

This research was funded by the US Army Tank-Automotive Command Research Development and Engineering Center (TARDEC) under contract DAAE07-97-C-X101. The views and opinions expressed in this paper are those of the author and do not reflect the policy or position of any agency of the United States Government.

8. REFERENCES

1. Blackwell, H. R., "Contrast thresholds of the human eye," *J. Opt. Soc.* 36: 624-43, 1943.
2. Cinlar, E., *Introduction to Stochastic Processes*, Prentice Hall, 1975.
3. D'Augustino, J., W. Lawson and D. Wilson, Concepts for search and detection model improvements, *Proceedings of the SPIE* 3063: 14-22, 1997.
4. Davison, A. C., and D. V.Hinkley, *Bootstrap Methods and Their Application,* New York: Cambridge University, 1997.
5. Johnson, A. E. and M. Hebert, "Surface matching for object recognition in complex three-dimensional scenes," *Image and Vision Computing* 16.9-10: 635-51, 1998.
6. Jonides, J. and H. Gleitman, "A conceptual category effect in visual search: O as a letter or digit," *Perception and Psychophysics* 12: 457-60, 1972.
7. Liu, Z. and D. Kersten, "2D observers for human 3D object recognition?" *Vision Research* 38.15-16: 2507-19, 1998.
8. Liu, Z., D. C. Knill, and D. Kersten, "Object classification for human and ideal observers," *Vision Research* 35.4: 549-68, 1995.
9. Mack, A. and I. Rock, *Inattentional Blindness*, Cambridge: MIT Press, 1998.
10. Marr, D., *Vision*, New York: W. H. Freeman & Co., 1982.
11. Moore, C. and P. Cavanagh, "Recovery of 3D volume from 2-tone images of novel objects," *Cognition* 67.1-2: 45-71, 1998.
12. Peli, E., "In search of a contrast metric: matching the perceived contrast of Gabor pathes at different phases and bandwidths," *Vision Research* 37.23: 3217-24, 1997.
13. Ratches, J. A. et al., "Night Vision Laboratory Static Performance Model for Thermal Viewing Systems," *R&D Technical Report ECOM-7043,* April, 1975.
14. Sun, J. Y. and P. Perona, "Preattentive perception of elementary three-dimensional shapes," *Vision Research* 36.16: 2515-29, 1996.
15. Tarr, M. J. and D. Kersten, "Why the visual recognition system might encode the effects of illumination," *Vision Research* 38.15-16: 2259-75, 1998.
16. Toet, A. et al., "A high-resolution image data set for testing search and detection models," *TNO-Report TM-98-A020*, Soesterberg, The Netherlands: TNO Human Factors Research Institute, 1998.
17. Ullman, S., *High-Level Vision: Object recognition and Visual Cognition,* Cambridge: MIT, 1996.
18. Washburn, A. R., *Search and Detection,* Military Operations Research Society of America, 1981.

COMPUTING SEARCH TIME IN VISUAL IMAGES USING THE FUZZY LOGIC APPROACH

Thomas J. Meitzler [a], Euijung Sohn [a], Harpreet Singh [b] and Abdelakrim Elgarhi [b]

[a] US Army Tank-automotive and Armaments Command
Research, Development and Engineering Center (TARDEC)
Warren, MI
E-mail: meitzlet@tacom.army.mil

[b] Wayne State University
Electrical and Computer Engineering Department
Detroit, MI
Email: hsingh@ece.eng.wayne.edu

1. ABSTRACT

The mean search time of observers looking for targets in visual scenes with clutter is computed using the Fuzzy Logic Approach (FLA). The FLA is presented by the authors as a robust method for the computation of search times and or probabilities of detection for signature management decisions in any part of the electromagnetic or acoustic spectrum. The Mamdani/Assilian and Sugeno models have been investigated and are compared. A 44 visual image data set from TNO is used to build and validate the fuzzy logic model for search time. The input parameters are the: local luminance, range, aspect, width, wavelet edge points and the single output is search time. The Mamdani/Assilian model gave predicted mean search times from data not used in the training set that had a 0.957 correlation to the field search times. The data set is reduced using a clustering method then modeled using the FLA and results are compared to experiment.

Keywords: Target acquisition, fuzzy logic, visual models

2. INTRODUCTION

It has been three decades since Prof. L. A. Zadeh first proposed fuzzy set theory (logic) [1]. Following Mamdani and Assilian's pioneering work in applying the fuzzy logic approach to a steam plant in 1974 [2], the FLA has been finding a rapidly growing number of applications. These applications include, transportation (subways, helicopters, elevators, traffic control, and air control for highway tunnels), automobiles (engines, brakes, transmission and cruise control systems), washing machines, dryers, refrigerators, vacuum cleaners, TVs, VCRs, video cameras, and other industries including steel, chemical, power generation, aerospace, medical diagnosis systems, information technology, decision support and data analysis [3, 4, 5, 6, 7].

Although fuzzy logic can encode expert knowledge directly and easily using rules with linguistic labels, it usually takes some time to design and adjust the membership functions, which quantitatively define these linguistic labels. Neural network learning techniques can, in some cases, automate this process and substantially reduce development time. To enable a system to deal with cognitive uncertainties in a manner more like humans, researchers have incorporated the concept of fuzzy logic into the neural network modeling approach. The integration of these two techniques yields the Neuro-Fuzzy Approach (NFA) [8]. The NFA has potential to capture the benefits of both the fuzzy and the neural network methods into a single model. Target acquisition models, based on the theory of signal detection or the emulation of human early vision, are not mature enough to robustly model, from a first principal approach without any laboratory calibration, the human detection of targets in cluttered scenes. This is because our awareness of the visual world is a result of the perception, not merely detection, of the spatio-temporal, spectra-photometric stimuli that is transmitted onto the photoreceptors on the retina [8]. The computational processes involved with perceptual vision can be considered as the process of linking generalized ideas, such as clutter or edge metrics [10], to retinal early vision data [9]. From a system theoretic point of view, perceptual vision involves the mapping of early vision data into one or more concepts, and then inferring a meaning of the data based on prior experience and knowledge. The authors think that the methods of fuzzy and neuro-fuzzy systems provide a robust alternative to complex models for predicting observed search times and detection probabilities for the vehicles in cluttered scenes that are typically modeled by defense department scientists. The fuzzy logic approaches have been used to calculate the search time of vehicles in different visual scenes within the commercially available MATLAB Fuzzy Logic Toolbox.

3. FUZZY MODELS AND WAVELETS

Fuzzy modeling of systems is an approach, which describes complex system behavior, based on fuzzy logic with fuzzy predicates using a descriptive language. Fuzzy logic models basically fall into two fundamentally different categories, which differ in their ability to represent different types of information. The first category includes linguistic models that are based on a collection of If-Then rules with vague predicates and use fuzzy reasoning. One of these reasoning mechanisms is based on the Mamdani and Assilian fuzzy inference method. Within this method, a scientist can design the membership functions manually and the output membership functions are continuous. The second method of fuzzy inference is based on the Takagi-Sugeno-Kang , or simply Sugeno's method. In the Sugeno method the membership functions are linear or constant.

Paper prepared for the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

For a review of these methods as applied to target acquisition modeling see [11,12].

The method of using wavelets to compute edge points, which are then used with fuzzy logic to compute the search time or the probability of detection, is derived from the elegant technique of Mallat and Zhong [15]. In [15] a derivation is made of 1- and 2-D wavelet transforms using a smoothing function, 0(x), that is a Gaussian. The integral of the function equals unity and the integral also converges to zero at infinity. We define the first- and second-order derivative of 0(x),

$$\psi^a(x) = \frac{d\theta(x)}{dx} \quad and \quad \psi^b(x) = \frac{d^2\theta(x)}{dx^2}. \qquad (1)$$

By definition the functions $\psi^a(x)$ and $\psi^b(x)$ can be considered as wavelets because their integral is equal to zero. The following subscript 's' will be denoted as the scale factor.

$$\varepsilon_s(x) = \frac{1}{s}\varepsilon\left[\frac{x}{s}\right] \qquad (2)$$

Following standard methods, the wavelet transform is calculated by convolving a dilated wavelet with the original signal. The wavelet transform of a function f(x) at the scale s and position x, calculated with respect to the wavelet $\psi^a(x)$, is defined in [15] as,

$$W_s^a f(x) = f * \psi_s^a(x). \qquad (3)$$

Similarly, the transform with respect to $\psi^b(x)$ is,

$$W_s^b f(x) = f * \psi_s^b(x). \qquad (4)$$

The above wavelet transforms are the first and second derivative of the signal smoothed at the scale or resolution level *s*. Substituting into (3) and (4) equation (2) for the 1-D case. Mallat then derives a 2-D expression for the wavelet transform of a function or image,

$$\begin{bmatrix} W_s^1 f(x,y) \\ W_s^2 f(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x}(f * \theta_s)(x,y) \\ \frac{\partial}{\partial y}(f * \theta_s)(x,y) \end{bmatrix} = s\vec{\nabla}(f * \theta_s)(x,y). \qquad (5)$$

The above wavelet transform definitions in (5) are important for a wavelet based clutter metric because they essentially define edge detectors that are used in the vision science community. For more discussion on this topic see ref. [16]. An implementation of eq. (5) in the program XWAVE was used to compute edge points.

4. IMPLEMENTATION

The Fuzzy Inference System (FIS) that models the relationships between the various input variables that affect the determination of the search time is done specifically for this dataset. The predicted search time for target detection can be determined with the FLA using input target metrics for the images shown below in Fig.'s 1 through 6. The input variables were; distance from the target to the observer (km), the aspect angle of the vehicle relative to the observer (deg), the target height (pixels) and the target area (pixels^2), target and the local background luminance (cd/m ^{2}), and the wavelet determined edge points of the scene as a measure of clutter. The one output parameter is the search time (secs). There were a total of 44 digitized color images along with the associated target and background metrics for the targets in each picture. 22 images are used for training and 22 are used for testing. Both the Mamdani and Sugeno type FIS methods are used and compared. The authors constructed the FIS's to predict search times using the MATLAB Fuzzy Logic Toolbox [13].

For convenience the algorithm for computing the wavelet edge points is summarized as follows:

- Read the input 256 X 256 element matrix which supports a discrete 2-D image f(x,y)
- Take the wavelet transform of the image (15) at a certain resolution level
- Set the threshold, here chosen as zero
- Determine the number of edge points in along with the number of pixels
- Find the edge density from the number of edge points divided by the total number of pixels
- Iterate s, the level of wavelet in the analysis
- Find the probability of detection (Pd) for the target in the scene.

Table I below lists the metrics used in the trials. The table entries, all except 'Edge points', were provided by Dr. Alex Toet of TNO. The entries are; target type number, distance from target to sensor, the absolute value of the sin of the aspect angle of the vehicle relative to the observer, the height of the target in pixels, the area of the target in pixels, the target luminance, the

darkest part of the target luminance, the surrounding area average luminance, edge points and the mean search time in seconds. The edge points were found using a wavelet program to compute the number of wavelet edge points over the whole image to give a measure of the clutter in the image.

Below in Fig. 7 is the Mamdani type FIS with the input parameters mentioned above and the search time as the single output. Fig. 8 is the firing array for the various membership functions using the Mandani approach.

Sample Visual Images

Courtesy of Dr. Alex Toet of TNO

Fig. 1

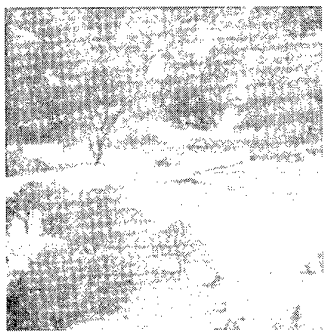


Fig. 2

Fig. 3

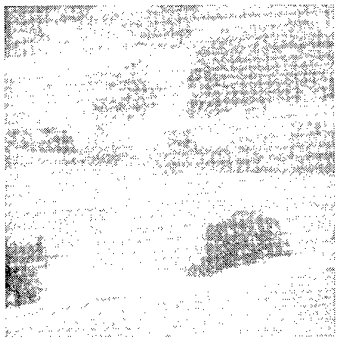


Fig. 4

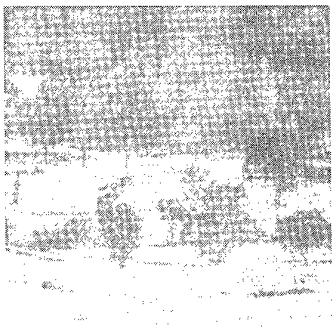


Fig. 5

Fig. 6

TABLE I Metrics for FIS construction

TARGET NO	distance	aspect	vert	area	target lum	Dark lum	Surround lum	Edgepts	SEARCH TIME
type	m	ass(sin)	pixels	*(pixels)*	scene	dark	grass	pts	search time(s)
1	4007	0.707	10	141	14	17	29	9571	14.6
1	2998	0.819	11	225	21	10	27	8927	15.2
2	3974	0.707	13	173	20	24	28	9138	12.4
3	5377	0.052	5	49	18	23	30	8970	29.8
2	1013	0.515	50	2708	19	5	34	8706	2.8
4	3052	0.000	11	100	12	18	30	8755	6.4
5	5188	0.407	9	76	18	23	28	9053	26.7
6	3679	0.122	10	96	12	20	26	8620	10.0
2	860	0.995	54	3425	9	1.5	40	8961	2.7
4	1951	0.848	16	332	15	11	27	8572	2.8
3	3992	0.788	11	154	20	19	26	9194	11.9
6	1041	0.743	24	1645	11	4	35	9074	2.5
7	2145	0.978	17	553	8	5	18	8280	3.7
3	1998	0.755	19	659	20	10	22	8739	8.1
2	4410	0.000	11	101	22	18	29	9404	12.4
1	2893	0.423	16	320	12	7	23	8670	2.5
5	1933	0.978	13	368	15	12	23	8606	4.8
1	1850	0.961	28	876	3	4	9	8464	2.8
8	1045	0.087	26	985	19	10	12	8613	12.3
2	1933	0.946	22	867	16	11	27	8376	2.8
7	4206	0.000	9	79	26	29	38	9506	15.1
1	5722	0.883	7	73	38	40	46	9044	25.6
4	4920	0.423	8	61	20	21	36	8618	12.1
6	4206	0.809	9	142	18	12	21	9152	8.0
5	2348	0.940	9	198	18	21	30	8504	5.5
1	3992	0.875	11	217	15	14	26	9078	7.8
9	4410	0.956	11	247	16	8	19	9397	9.6
8	2321	0.829	15	458	22	21	47	8365	5.1
5	3661	0.755	9	84	17	25	23	8807	7.5
3	3670	0.000	13	192	14	15	27	8483	6.1
7	1671	1.000	19	893	15	13	31	8959	3.5
4	4345	0.809	8	63	15	12	20	9021	12.3
2	3662	0.574	10	203	26	25	44	8702	5.4
5	633	0.707	50	4403	20	5	39	8741	2.5
3	492	0.070	57	3045	20	16	23	8992	2.2
4	1497	0.777	16	560	10	7	20	9014	5.8
5	1041	0.999	33	1613	17	5	32	8486	2.6
1	2891	0.985	19	486	12	12	35	9021	12.1
7	5147	0.934	5	81	18	27	34	9075	34.9
6	1648	0.588	18	648	23	7	37	9070	2.7
8	948	0.731	35	1463	18	5	38	8790	3.7
7	3662	0.407	12	188	19	25	39	8524	5.8
6	2900	0.000	17	340	20	10	49	8791	4.1
2	5136	0.000	10	79	25	16	27	8941	10.6

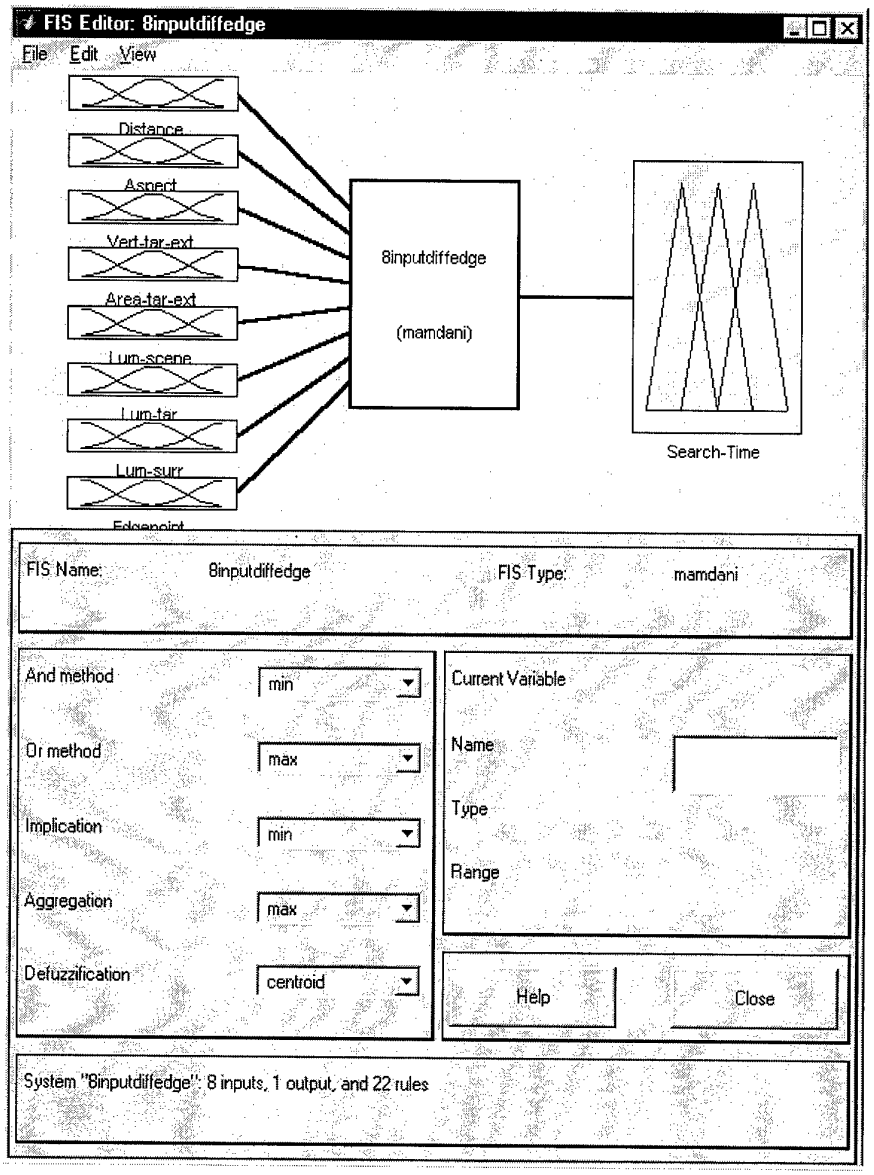


Fig. 7. Mandami Fuzzy Logic Identification System for computing visual search times

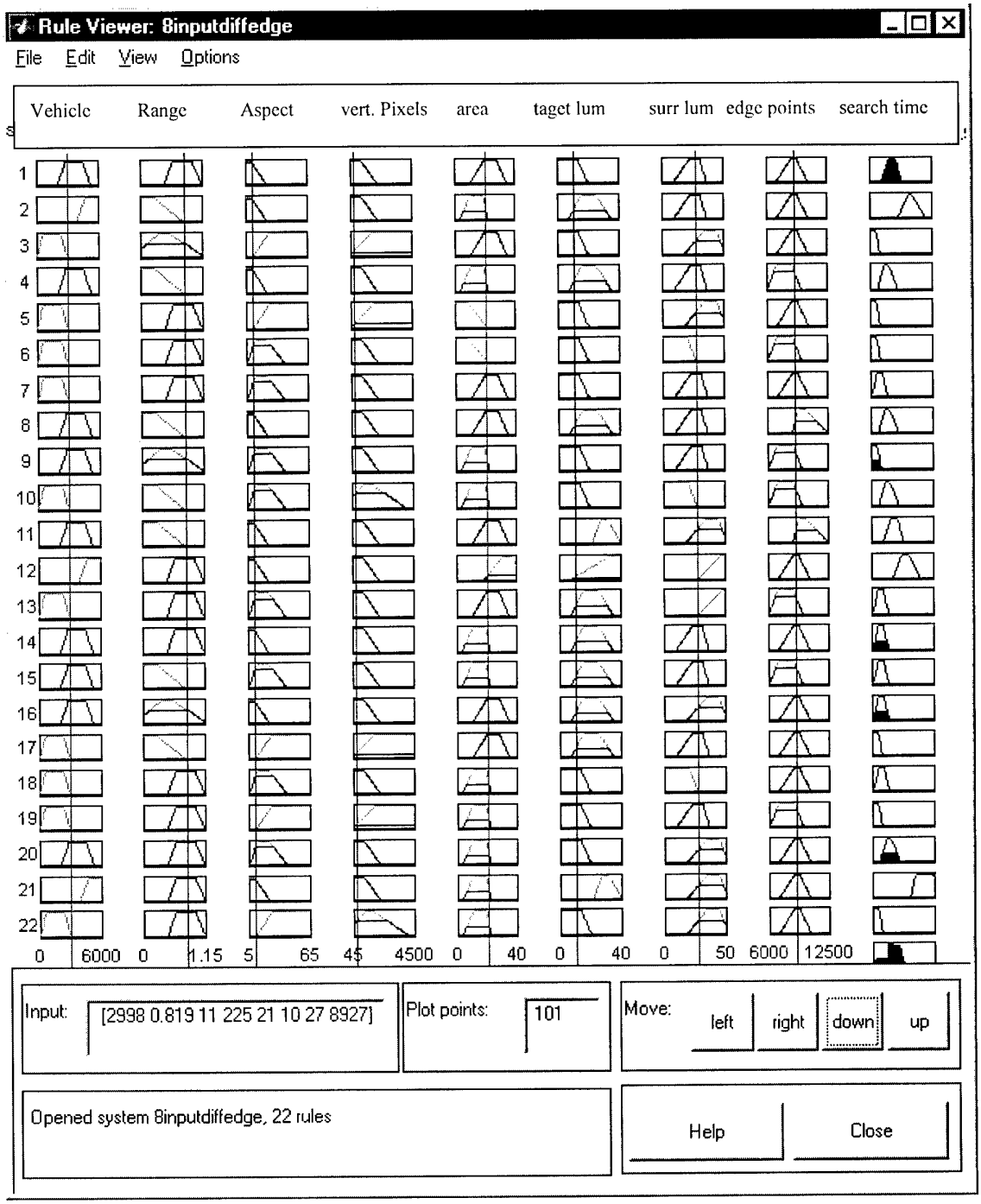


Fig. 8 Firing diagrams for the Mamdani FIS to predict search times

5. RESULTS

Fig. 9 shows the correlation of laboratory search times to FLA predicted search times using the Mamdani approach with membership functions we designed and achieved a 0.957 correlation of model predicted search times to experimental search times. Fig. 10 is the output of the ANFIS model of the data, which gave a 0.60 correlation to the data. We also tried using the Mamdani FIS , with the 0.957 correlation to experiment, on another data set of visual imagery [14]. The FIS from one data set can be used to model another data set, if and only if, the metrics used to describe the various data sets are similar.

These results are indicative of the power of using the FLA to model highly complex data, for which there would be many interrelated equations if one tried to model the detection problem in the conventional standard algorithm based method.

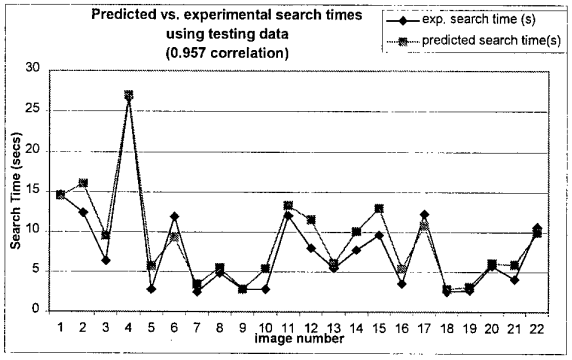


Fig. 9 Graph of search times from Mamdani FLA model and the laboratory

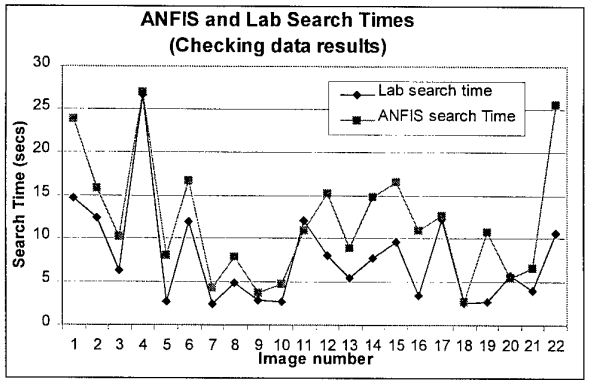


Fig. 10 Chart showing the comparison of experimental search times to ANFIS FLA predicted search times

Clustering was also used to model the visual metrics and responses. For a large dataset, it will be desirable to reduce the number of input vectors to a small number to reduce the number of rules and membership functions that need to be constructed. Clustering was used to obtain the means of the 7 input vectors. The center of the clusters was used in the construction of the membership functions. The correlation results are shown below in Table 2. Clusters were made of 15, 18 and 20 data points. The FLA with clustering was used to predict search time for the 22 points not used in obtaining the clusters and for the entire data set of 44 images.

TABLE 2 System Evaluation Using Cluster Centers

Cluster	Correlation for 22 points which are not used for clustering.	Correlation for 44 points
fcm15	0.83	0.85
f18	0.75	0.82
fc20	0.82	0.88

It is expected that increasing the number of cluster centers and the number of rules will improve the correlation. This is not the case when the number of clusters was increased from 15 to 18. The reason for this is due to the random operations used in clusters' center calculations. In other words, if we started from another clusters' center we may get better correlation. We used the cluster centers as the centers of membership functions, but chose initial values for the width. We can then tune the width manually to increase the correlation. It is clear that there needs to be an objective algorithm or technique to tune the width of the membership functions as ANFIS does. Below in Fig. 11 is a snapshot of the result of clustering the input variable distance over 15 cluster means.

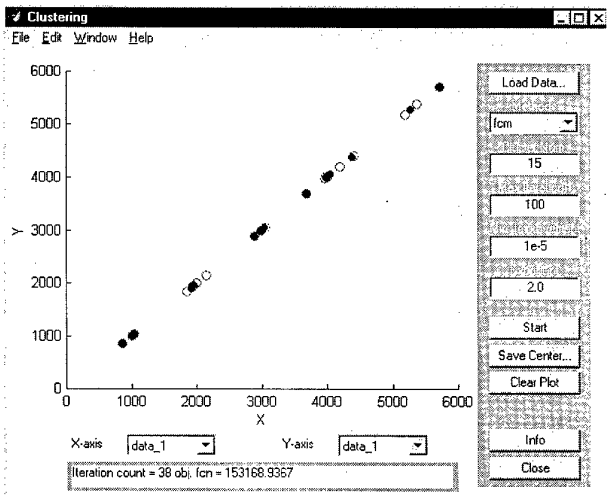


Fig. 11 Clustering results using 15 clusters

6. CONCLUSION

In conclusion, the FLA yields very satisfactorily results. 0.97 correlation of laboratory or field data to model predicted data, and requires a fraction of the effort that goes into traditional algorithm based techniques of modeling target acquisition probabilities and search times. We expect that the fuzzy modeling approach could be used in the existing statistical decision theory modules of target acquisition models for any spectral regime.

Two fuzzy models have been used: namely the Mamdani and Sugeno models. This application of the FLA involved pictures, metrics, and experimental search times of images in the visual band. Future work will involve the application of the FLA to predict the Pd's of ***moving*** targets in visual and infrared cluttered scenes for military and commercial applications. Clustering of the input data was explored as a means to reduce the number of input vectors and membership functions. For large data sets, a saving of computational time and effort should be realized using this approach. The membership functions can be designed using experimental Pd's or search times collected in the TARDEC Visual Perception Laboratory (VPL). The TARDEC VPL is being used in a collaborative R&D project with auto companies on vehicle conspicuity.

ACKNOWLEDGMENTS

The authors extend thanks to Dr. Alex Toet of the TNO Human Factors Research Institute of the Netherlands for providing the visual images for study and conspicuity metrics used in their vision research and for the assistance of Mr . Deok Nam from the E.C.E. Dept. Wayne State University.

7. REFERENCES

[1] L. Zadeh, "Fuzzy Sets", Information and Control", 8, pp. 338-353, 1965.

[2] E. Mamdani and S. Assilian, "Applications of fuzzy algorithms for control of simple dynamic plant", Proc. Inst. Elec. Eng., Vol. 121, pp. 1585-1588, 1974.

[3] T. Munakata, and Y. Jani, "Fuzzy Systems: An Overview", Commun., ACM, Vol. 37, No. 3, pp. 69-76, Mar. 1994.

[4] R. Mizoguchi, and H. Motoda (eds.), "AI in Japan: Expert Systems Research in Japan", IEEE Expert, pp. 14-23, Aug. 1995.

[5] E. Cox, *The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems*, AP Professional, 1994.

[6] D. G. Schwartz, G. J. Klir, H. W. Lewis, and Y. Ezawa, "Applications of Fuzzy Sets and Approximate Reasoning", IEEE Proc., Vol. 82, No. 4, pp. 482-498, Apr. 1994.

[7] T. Terano, K. Asai, and M. Sugeno, *Fuzzy Systems and its Applications*, AP Professional, 1992.

[8] J-S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System", IEEE Trans. Sys., Man, and Cyber., Vol. 23, No. 3, pp. 665-684, May/Jun. 1993.

[9] M. Gupta and G. Knopf, "Fuzzy Logic in Vision Perception", SPIE Vol. 1826, Robots and Computer Vision XI, pp. 300-276, 1992.

[10] S.R. Rotman, E.S. Gordon, and M.L. Kowalczyk, "Modeling human search and target acquisition performance:III. Target detection in the presence of obscurants," Optical eng., Vol. 30, No.6, June 1991.

[11] T.J. Meitzler, "Modern Approaches to the Computation of the Probability of Target Detection in Cluttered Environments", Ph.D. Thesis, Wayne State University, Dec. 1995.

[12]T. Meitzler, L. Arefeh, H. Singh, and G. Gerhart, " The fuzzy logic approach to computing the probability of target detection in cluttered environments," Optical Engineering, vol. 35 No. 12, December 1996, pp. 3623-3636.

[13] *Fuzzy Logic Toolbox*, for use with the MATLAB, the Math Works Inc., Jan. 1995.

[14] Meitzler, T, Singh, H., Arefeh, L., Sohn, E., and Gerhart, G., "Predicting the Probability of target detection in static infrared and visual scenes using the fuzzy logic approach," Opt. Eng., Vol. 37 (1), Jan. 1998.

[15] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No.7, pp. 710-732, (1992).

[16] Meitzler, T., Karlsen, R., Gerhart, G., Sohn, E., and Singh, H., "Wavelet transforms of cluttered images and their application to computing the probability of detection," Optical Engineering, 35(10), pp. 3019-3025, Oct. 1996.

THE SOURCES OF VARIABILITY IN THE SEARCH PROCESS

K. Cooke
BAe SRC
FPC267
Filton, Bristol
BS34 7QW
England.
E-mail: kevin.cooke@src.bae.co.uk

1. SUMMARY

Modelling of camouflage concealment and detection, needs to consider the terrain in which a target will appear. A variety of capabilities for evaluating target signatures through to the human response now exists. The modelling approach adopted at British Aerospace research centre for many years has been a statistical one. Although image analysis techniques have been explored it has been cost effective for our purposes to stay with the statistical model ORACLE.[1] A complex problem in modelling human visual performance is to find an adequate relationship between recognition thresholds across the visual field and simple target descriptions. The ORACLE model represents recognition as the resolution of a fraction of the target perimeter. The fractional perimeter concept has been applied further to representing average observer performance in structured scenes. Modelling of the SEARCH 2 data[2] was found to need a similar distribution of fractional perimeter values to a previous UK field trial. The use of a statistical lobe model for analysing search and recognition can be supported for generic information.

Keywords ORACLE, visual lobes, recognition, identification, search variables, field data, peripheral, search time.

2. INTRODUCTION

The issues raised when considering target acquisition can be divided into the visual and the cognitive components. From a physical design stance the cognitive issues are of less interest than the visual issues because physical design can impose specific visual limits on system performance. The development of ORACLE in British Aerospace has been focused on the products of the company and how to optimise visual performance. Around 1983 a decision was made not to attempt to model performance in specific scenes. At the time we were studying eye movement patterns and we were synthetically altering clutter levels in scenes. Computer power at the time made the analysis of the data slow compared to today but we also realised that the scenes we were analysing took a long time to generate and our customer base was mostly interested in the relative performance of one system design against another. Our belief was that effort spent on trying to unravel the complex interactions of clutter, strategy, training philosophies, etc, for inclusion in the ORACLE vision modelling approach was not cost effective. We had at the time developed modelling primarily for ground to air acquisition tasks and relatively uncluttered or plain backgrounds were appropriate. We chose to build a search model based on the detection visual lobe and its relationship to search in plain fields of view. The search modelling has remained unchanged in ORACLE for the last 16 years. We rationalised cluttered scene search to the use of visual lobes based on recognition or identification criteria. Search is a combination of peripheral cueing and foveal interrogation. A target in a structured scene may be surrounded by confusable objects and the task of the observer is to use peripheral discrimination for guiding fixation. Our belief was that at any one time an observer is using what could be thought of as a multitude of visual lobes operating from pure energy detection through to discrimination and that the largest signal to the visual system provides the cue for the next point of fixation. A far peripheral target that is highly detectable may provide the same strength of cue for fixation as a near but confusable object. We have attempted to model the variance experienced in structured scene search by using a set of visual lobes to allow for observer, target, and background statistics. In this way we hope to have encompassed the range of statistical variance that has been measured in field or laboratory trials. This simplification has avoided detailed consideration of the influence of strategy and other cognitive elements of the search task for purely pragmatic reasons. Our approach has excluded specifying cognitive processes that occur when an observer is searching in a specific scene.

In this paper I would like to provide an indication of some areas we have investigated for lobe modelling and its relevance to sensor design. Also to provide a feel for the sensitivity of the model and why we have addressed some issues and not others. As long as models have the sensitivity to deal with the parameters that influence visual performance in sensor/display design then advances along the cognitive dimension of the search task will hopefully enable further improvements in total system performance prediction.

3. DEVELOPMENT STRATEGY

We have attempted to make the models 'user friendly' to non vision scientists and to engineers with access to standard computer systems. As the outside world poses system limitations the early work was devoted to providing measures of target and atmospheric characteristics. Our philosophy was to calculate separately parameters of the displayed information reaching the cornea and then apply a generic model of human vision which would be applicable to any sensor/display system.

I believe that the processes contained within models for representing human vision are less important than the model output. If we had test data sets agreed between modellers the method of modelling them could be independently derived, and I applaud the aim of this meeting to work with a common data set. Some models have aimed at predicting the threshold surface for size, contrast, and retinal position and several early data sets from Blackwell[3] have been well used for test purposes. We also know that tunnel vision makes search performance poorer, thus emphasising the importance of peripheral vision. We have spent

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

the bulk of our time trying to get the local retinal performance modelling accurate by concentrating on the physical attributes of a stimulus such as size, contrast, colour, motion, and image quality across the whole visual field. These are relevant to lobe evaluation. It is only when we have the lobe well represented that we can begin to look at the nature of successive fixations in further depth. Hence the still very rudimentary nature of our search model[2]. Our search model evaluates the target against its immediate background but does not analyze the visual information contained in the rest of the FoV and does not include specific observer strategies in the equations. The need to account for scene interactions becomes a necessary part of specific scene analysis but is less crucial for the generic scenario. The variance is great as is highlighted in the SEARCH 2 data set as well as many others.

There are now a variety of direct image processing models which can process the whole scene rather than just the local target to background contrasts. As an industrial group we are still debating the cost effectiveness of attempting to do this using the ORACLE concepts. We still ask the question whether by adding several orders of complexity would we now meet customer requirements to a significantly better degree? We have continued to devote our efforts to the more easily quantifiable processes leaving the cognitive variances of stress, training, learning, etc as poorly quantified components of our observers.

The calibration of the sensor from which digital data are recorded should be well understood if full image analysis is to be correctly utilised. The range of intensity/colour levels achievable when displaying images is restricted compared to the real world and some effort would need to be devoted to creating equivalent visibility images. Calculation of visual performance in extreme viewing scenarios has been a crucial point in the application of our model. Atmospheric effects on a field evaluation or when simulating a future event are not very predictable and are not controllable, and the resulting effects on contrast and image sharpness can be large.

The rest of this paper is a review of issues to do with applying lobes to search tasks concentrating on some of the fundamental components of our vision model and the associated data sources that have proved useful for calculating visual lobes for application to acquisition.

There are still many questions for which there are insufficient data to establish model validation. For example can we recognise objects with rod vision or do we rely solely on cone vision. The interface between rods and cones is not easily bridged and pure rod or pure cone data are sparse.

4. RECOGNITION

Representing recognition in a statistical model involves simplification of the target signature. The approach in ORACLE is similar to others such as the Johnson criterion which is essentially an equivalent resolution, and to the equivalent disc concept proposed by van Meeteren[4]. Immediate limitations of the above are that the Johnson resolution[5] bar patterns operate only in one dimension and the equivalent disc fails at resolution limits. Our approach is to define the resolution requirements to be a fraction of the vertical and horizontal dimensions of the target. This alone does not provide the necessary behaviour to deal with all size-contrast regimes and the second component of the model, more recently introduced, requires a minimum number of elements of the target object to be separately resolved for recognition to be achieved.

This aspect of our modelling has often required the greatest explanation to potential users of the model as the implementation of the equation to represent recognition reduces essentially to resolving features in a statistical fashion but does not specify exactly which features of a target are represented.

The support for our approach is derived from a series of experiments which provide the calibration and test data for establishing the fractional dimensions of the target perimeters that correlated to recognition and identification performance. The tasks, from energy detection through to detailed object discrimination, are viewed as a continuum with the resolution required for the detail of the target object decreasing in absolute size as the task shifts from detection to identification. Our earliest study [2] was to measure size thresholds for recognition and identification of a set of images of tanks, trucks, bushes and buildings. The observer's task was firstly recognition, which was placing the stimuli into the above categories, and secondly identification which required specification of the type of vehicle. The sizes of the targets are shown in Figure 1. The perimeter sizes are quoted in mrads perimeter. All the stimuli were objects of an average luminance contrast of -0.26 and from this the ratio of the detection size to the recognition size or identification size provides the fractional perimeter value.

The generality of the approach was tested further with alphanumeric characters. Experimental thresholds for recognising numbers and letters were measured (Figure 2). Prior experiments by Bowler[6] had found that the alphabet had threshold sizes with a distribution which centred around the Landolt C.

The Landolt C is characterised by a gap one fifth of the height and a stroke width one fifth of the height. The model uses a median fractional perimeter value of 0.2 for modelling recognition of alphanumeric characters.

Limiting the relationship to just a fraction of the perimeter fails with sampled stimuli as the target object can be large and the detailed information may be lost. Studies of sampling on recognition[7] reveal a linear relationship between required target size for recognition and sample size. The number of samples over the linear range was a constant for a given task.

The model therefore required a second component that limited recognition probabilities for conditions with large stimuli at high luminance contrast and with very coarse sampling. For a given task level we have included a function that models a required number of elements of the target object that must be resolvable before recognition can be achieved. If the sampling of the displayed information is not the limiting parameter then the limit is imposed by the optical spread function and receptor sampling of the eye. Data from legibility experiments from Ginsburg[8], Tomoszek[9] and Bowler[6] (Figures 2 and 3) are shown, which provide evidence of the limiting sizes below which observers are unable to discriminate characters. Figure 4 shows the data from the sampling experiment with the trend for increased target size with sample size.

Having introduced the above concepts to the model and using the standard search model we could then establish a distribution of fractional perimeter values for search tasks. Figure 5 shows the fractional perimeter distribution for laboratory measured foveal recognition and identification as shown earlier in Figure 1. Figure 6 results from an analysis of the SEARCH 2 data [7] set and shows the distribution of fractional perimeter values which were arrived at by best fit between modelled median time and experimental median times for individual scenes and sets the required task level for modelling average performance of

observers for the scenes. A previous exercise with a UK field trial revealed a very similar distribution. Our experience has shown that lobes that are representative of tasks from recognition to identification cover the bulk of structured scene search tasks. We define recognition here as putting objects into general classes of truck, AFV, bush or building, and identification as naming the type of vehicle.

5. PERIPHERAL VISUAL FIELD THRESHOLD CONTRAST DATA

The threshold contrast trends for peripheral viewing show a consistency between cone receptor density and small target thresholds. Detection has been measured peripherally by Taylor[10] for single glimpse viewing using an exposure duration of 0.25 seconds. The scaling of thresholds for small targets provides data related to acuity studies. Modelling can represent peripheral visual performance by using a simple relationship to receptor density and the receptive field or cone sampling aperture in combination with the optical effects on the MTF to provide a scaling of detection with eccentricity.

6. ISSUES IN MODELLING SEARCH FROM LOCAL VISUAL PERFORMANCE

The lobe will not model much more than the pure visual components of the search process although we have found that next fixation can be calculated with a good accuracy for moderate clutter scenes where glimpse to glimpse patterns are governed by peripheral signal strength for a single confusable item. Where strategic choices are needed then the lobe model is not sufficient. We know that using technical training manuals as guides to how observers search in order to develop strategy in models is not ideal, as we have found that some military observers will recite the manual when we ask how they search, but when measured using eye movement analysis the patterns used were different. Foreground, midground, far ground priority did not feature strongly in their eye movement patterns.

6.1. Closing Range

For some tasks simple modelling is very appropriate. Our early studies into the detection of closing range aircraft found that the rate of target growth was the dominant component. Where the scene is relatively uncluttered and the task is the detection of an energy difference then the growth of the intensity of signal for a rapidly closing aircraft is dominated by the increase in signal intensity, and simple lobe models are very effective.

6.2. Probability or signal space

Does it make sense to model search in probability space? We have had some success in predicting the next fixation by modelling the level of signal above threshold in combination with an aspect ratio contrast and size difference contrast. These were for contrived scenes where objects were stretched or brightened to make signal strength in one dimension noticeably different, and a high success rate was achieved in predicting the next fixation.

6.3. Search time statistics

We assume the observer is dedicated to the search task and that time sharing is not a factor. There are many ways of analysing search data but the method chosen must be consistent with the modelling approach to be a fair test. The statistics of search times are not normally distributed unless transformed into log or alternative space, therefore mean times and 50% cumulative probability times provide different values. Reaction time, particularly in an easy task, needs to be removed from the analysis unless it is included as part of the modelling consideration.

6.4. Signal to noise

There is a need to consider experimental methods and their effects on visual performance. For example the difference between forced and free choice on an observer's threshold is pronounced. The decision process may have an attentional component but we have very limited evidence to suggest whether signal to noise threshold criteria are maintained equally across the retina. A similar problem applies to target size and whether the process is as sensitive with small high contrast targets compared to large low contrast targets. Factors of 2.5 to 6 above Blackwell's 1946 [3] threshold have been suggested. Typical minimum contrast thresholds in a practical task are rarely less than 1%.

6.5. Temporal variation

Every threshold measurement is subject to a variance and specific studies have examined the change of threshold with time. The inclusion of these variances is part of the search process. We allow for a small variance between glimpses but a larger one over minutes.

6.6. Area under visual lobe to search

The use of plain fields of view has provided control over target signature, and results from experiments have generated both lobes and search data, from which a basic search model has been developed. We can establish a high correlation between area under the lobe and the rate of accumulation. We have also needed to allow for within and between observer variability to represent the search process.

6.7. Foveal Scotoma in a search task

The loss of equal volumes of lobe probabilities during a search task do not lead to equal search performance if the fovea is involved. This may be due to a foveal dependence of accommodation mechanisms, or due to fixation times taking longer due to loss of high acuity vision. The relationship of the lobe to search is not simple because the loss of the fovea considerably limits the speed of target acquisition. Measured eye movements do produce an increase in the mean fixation duration but not significant enough to account for the change of search probability as shown in Figure 8. The measures of saccadic amplitude are not significantly different under the influence of a simulated scotoma.

Table I. Comparative data for peripherally designated non-targets and targets.

Non targets Eccentricity (degrees)	Aspect ratio	perimeter length (mrads)	Luminance Contrast
0-1	1.4	35.8	1.5
1-2	1.4	38.1	1.8
2-3	1.7	32.0	1.5
3-4	1.4	31.0	1.4
4-5	1.2	32.0	1.3
5-6	1.1	33.4	1.6
6-7	1.1	34.8	1.8
7-8	1.2	36.2	2.1
> 8	1.3	40.9	2.5
Average	1.3	34.1	1.7
Targets	1.6	22.7	1.3

6.8. Cluttered scenes require recognition lobes

The useful visual lobe in a search task cannot be easily measured and may not exist. Fixation cueing may be identification foveally, recognition near periphery and detection in far periphery. The modelled lobe is in concept the average performance lobe for carrying out search. Ideally we could calculate target signal strength parameters related to glimpse position. Measurement of recognition lobes for military targets in structured terrain involves discrimination of the vehicles from equivalent clutter. An analysis of the objects falsely designated as targets showed that they had near constant aspect ratio, larger perimeter length and higher contrast than the average target object and so were of greater signal strength.

Experiments to establish discrimination capability of peripheral vision using military vehicles as targets were conducted using images captured by video camera from a terrain model board[2]. The model was 300:1 scale and contained villages and rural scenery. Static images, some containing targets were displayed to the observer for 0.33 seconds while the observer fixated centrally. Targets appeared in the scene at eccentricities of up to 9 degrees. Figure 9 shows the resultant average visual lobe for recognition of targets taken from 515 scenes containing targets.

An additional analysis of the confusable objects that were falsely designated as targets provided useful information on the likely cues that observers were using in designating an object as a target. Table 1 is a list of the aspect ratio, perimeter length and luminance contrast of the objects. The table includes both the non-targets and the targets.

Sizes and contrasts of the non-target objects meant they were more detectable than the targets contained in the same scenes. The above table shows non targets were generally of longer perimeter length and greater luminance contrast than the targets but with smaller aspect ratios.

6.9. Glimpse time

Do we agree that the average glimpse, which consists of fixation and saccade has a duration of 0.33 seconds. We know the process is saccadic. Several data sources support the average glimpse time of 0.33 seconds. The glimpse patterns of individuals are different and a random glimpse statistic becomes indistinguishable from the glimpse fixations of multiple observers.

6.10. Colour

We have not successfully developed a colour model which directly uses the cone-integrated spectral signals in a manner that can be related to proposed theoretical processes in the visual system. Our most recent attempt at implementing the model of DeValois and DeValois[11] was unsuccessful at modelling threshold and suprathreshold search processes. We could not model a visual signal level that was adequate across a variety of luminance conditions. Our own tests on the combination of colour and luminance data have led us to use the CIE u'v' space for calculation of colour difference. Our calculation of colour contrast is

$$C_c = \sqrt{(u'_T - u'_B)^2 + (v'_T - v'_B)^2}$$

where C_c is the colour contrast, and u' ,v' are CIE co-ordinates for the Target (T) and background (B). This is substituted into the luminance equation and a constant included for relating colour signal to measured performance.

To account for target motion effects on the highest acuity channel, performance is modelled as a degradation with increasing velocity. The ability to track a moving target is important and the loss of efficiency at tracking with increasing velocity provides our modelling approach. The modelling is based on the change of threshold acuity with velocity using data from Ludvigh and Miller[12], and Barber[13] A function representing efficiency of foveal fixation and therefore an estimate of duration of exposure on a retinal location, will obey Bloch's law. No effect is apparent below 0.125 degs/sec which is the average eye motion due to tremor and drift components of eye movement. Degradation is not large at velocities below 30 degs/sec where smooth eye tracking is achieved.

6.11. Target number

A study of multiple targets in a scene showed that overall detection probability decreased with a large number of targets. Target numbers of 1, 6, 12, and 30 per scene and with densities of 0.75, 4, 8 and 20 vehicles per square kilometre showed that the final probability of finding all targets decreased although the probability of finding the first target was significantly higher.

Figure 11 shows target acquisition probability against time for varying target density and Figure 12 for the progressive detection of targets.

6.12. Eye movements in search studies with plain Fields of View (FoV).

Eye movements of observers show a full coverage of field of view including the sky which we believe was for orientation purposes. The following conclusions were drawn by Bell [14] from eye movements studies during a search task.

1. Observers tended to neglect some areas of the FoV, most consistently the 1 degree at the edge.

2. The number of wasted glimpses decreases with increasing field of view size. (16.8%in a 5 degree FoV, 12.2 in 10 degree FoV, and 8.2 % in a 20 degree FoV) most were within 1 degree of the FoV edge.
3. The shape of FoV had no influence on fixation distributions.
4. Cross wires in the FoV had the effect of increasing the number of wasted fixations and altering the strategies that observers used to search.
5. Search strategies can be modelled as random although an individual directs his search pattern in an orderly manner.

As the size of the field of view increased observers tended to organise their search strategies in a more orderly manner. There was a consistent reduction in cumulative probability with time for increasing size FoVs. There is a considerable variation within performance of observers even in plain fields of view. Interfixation distance decreased significantly from a length of 3.3 degs for a 20 deg FoV, to 2.3 for 10 deg FoV, and 1.6 degs for a 5 deg FoV. These experiments involved the introduction of the target at an unknown time interval to the observer. The search strategies may have had an impact dependent on introduction time.

Can we carry over any of the properties of an empty field search model to a structured scene?

1. Glimpse frequency is consistent.
2. There is randomness in a population of observers in a plain field and similar effects did occur in structured scenes, although this is less likely to be a major influence on search time.

6.13. Search slewing influences on fixation

Holman (1985)[15] showed that observers put under a time constraint achieved a higher glimpse rate. If the field of view is slewing under the observer's control there is a concentration of fixations on the centre of the display. If the sensor is moving under independent control then fixations were towards the leading edge, with confirmatory saccades tracking back to cross check confusable objects in the scene.

7. CONCLUSIONS

We have not yet explored sufficiently the cognitive components of search and target acquisition. Search tasks in a military scenario can involve vigilance, fear, stress, expectation, strategy, pre cueing, etc. It is easier to model those data which are simpler to measure as we can define them accurately. We need to measure our data test sets comprehensively so that as we progress from the simple to the complex cognitive task, we can test the intermediate stages. All laboratory studies by their very nature limit the variables to a fixed number, and impose control beyond a level experienced in natural practice. The use of models derived from such behaviour must be suspect for the practical environment, hence the necessity for a degree of field validation. Our experience of terrain model simulation is that short range targets are often missed. We have attributed this to observer expectation. The same seems apparent in field trials but whether it happens operationally we do not know. In some of the studies where short range targets were not consistently detected[16], eye movement studies showed that observers fixated on the target several times before eventually responding to a positive detection. The fixations prior to designation were substantially longer than others. Observers briefed to expect short range targets were more likely to detect them.

The visual lobe is an essential building block for the search process and to use it requires the capability to model foveal and peripheral vision. We can show that next fixation is predictable in some scenes from a series of target related measures. We must also consider the observer's adaptation state which requires knowing about the status of the observer before search commences; such data we have found in the past to be scarce Therefore dependent on the task set to the modeller there is a choice of statistical versus image based modelling. Statistical models provide general performance and cannot be applied to anything other than a statistical selection of scenes. Use of a single real scene is an inadequate approach for camouflage assessment and a multitude of background types and target ranges and atmospheric conditions must be considered. With enough scenes specific image analysis will contain the variance which we have attempted to put into the statistical model. There is a place for both statistical modelling and image modelling, both rely on greater understanding of human visual processes and the choice is probably dependent on cost effectiveness with both reaching approximate answers. No vision model can provide the perfect answer as all our models are built to be an approximation of the truth.

8. REFERENCES

1. Cooke, K.J., *The ORACLE Handbook*, British Aerospace Sowerby Research Centre, BAe Report JS12020, 1992.
2. Toet, A., Bijl, P., Kooi, F.L., and Valeton, J.M., *A high resolution image data set for testing search and detection models,* TNO-HFRI Report TM-98-A020, 1998.
3. Blackwell, H.R., "Contrast thresholds of the human eye", *J.Opt.Soc Am* 36, pp. 624--643, 1946
4. Van Meeteren, A. , "Characterization of task performance with viewing instruments", In: *Vision models for target detection and recognition*, Edited by Eli Peli, World Scientific publishing Co., 1995.
5. Johnson, J., *Image Intensifier Symposium,* Fort Belvoir, VA, AD220160, 1958.
6. Bowler, Y.M., *Recognition of simplified characters in support of ORACLE modelling, BAe Report* JS11104, 1988.
7. Stanley, P.A., Cooke, K.J., and Davies, A.K., *Modelling the effects of array motion on target recognition with sampled imagery*, BAe Report JS13700, 1997.
8. Tomoszek, A., *Size versus contrast thresholds for the detection of numeric symbols,* BAe Report JS12603, 1993.
9. Ginsburg A. P., *Visual information based on spatial filters constrained by biological data,* Phd Thesis, University of Cambridge, 1978.
10. Taylor, J.H., *Contrast Thresholds as a function of retinal position and target size for the light-adapted eye*, Scripps Institute of Oceanography, Report 61-10, 1961.
11. De Valois, R.L. and De Valois, K.K., "A multi stage colour model", *Vision Research*, 33, pp. 1053-1065, 1993.

12. Ludvigh, E. and Miller, J.W., "Study of visual acuity during the ocular pursuit of moving test objects. I. Introduction", *J. Opt.Soc Am,* 48, pp. 799-802, 1958.

13. Barber, J.L., *The spatial organisation of movement-detection mechanisms of human vision,* PhD Thesis, Submitted to the University of London Imperial College of Science and Technology, 1980.

14. Bell, J.B., *Visual lobes for complex scenes,* BAe Report BT13006, 1982.

15. Holman, L.K.B., *Visual search strategy predictions : the use of aspect ratio as a cue,* BAe Report BT12565, 1981.

16. Carr, K.T., *An investigation into the acquisition of short range targets in visual search, BAe Report* JS10693, 1986.

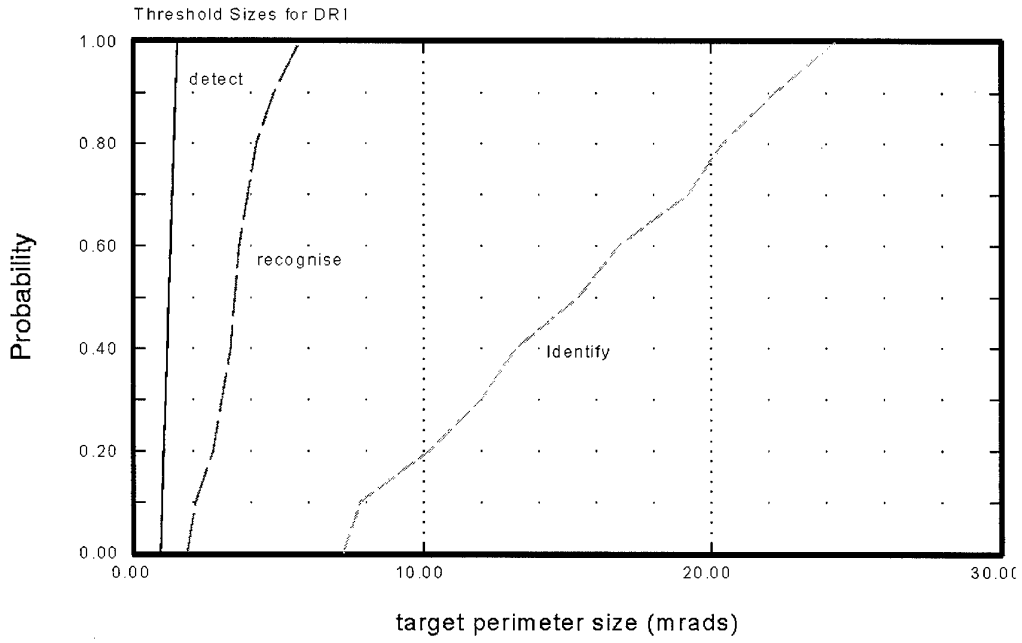


Figure 1. Detection Recognition and Identification probabilities against target size.

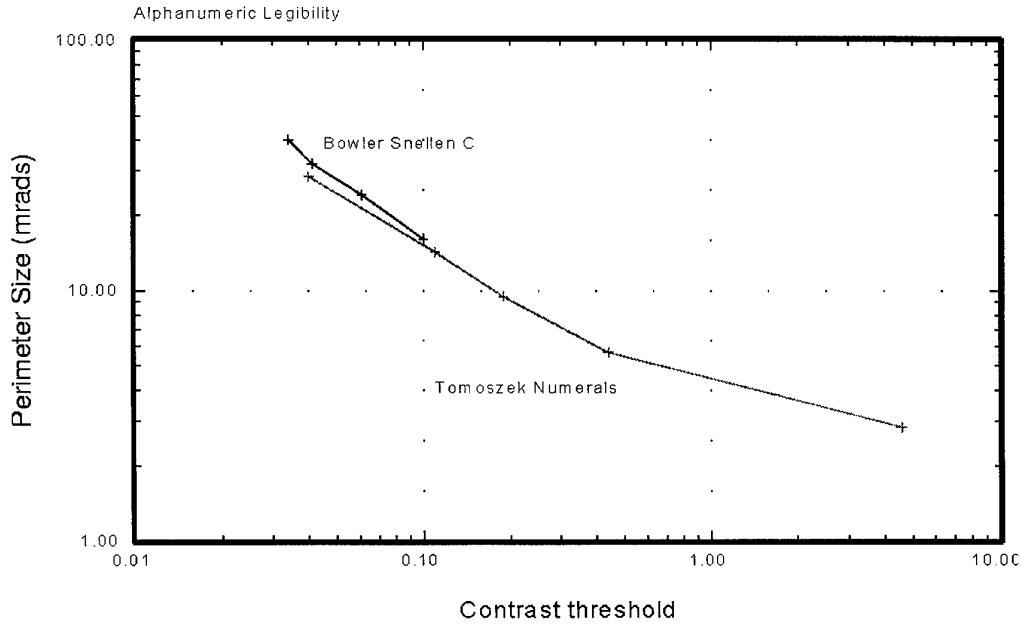


Figure 2. Alphanumeric character recognition experimental threshold data taken from Bowler[2] and Tomoszek[3].

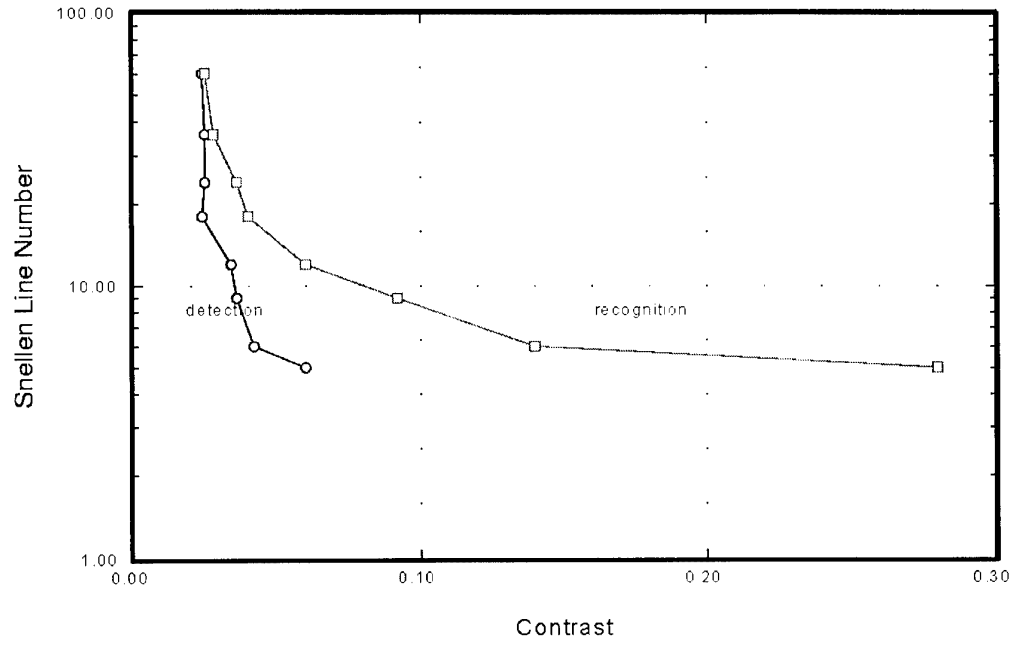


Figure 3. Ginsburg[9] character recognition data for variable size and contrast.

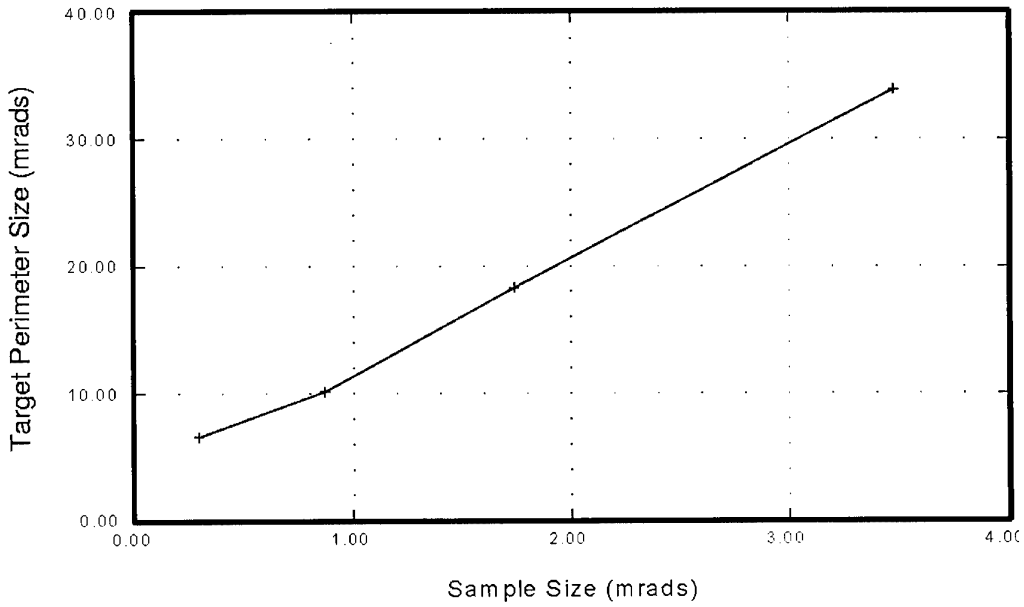


Figure 4. Recognition threshold sizes dependent on sample size.

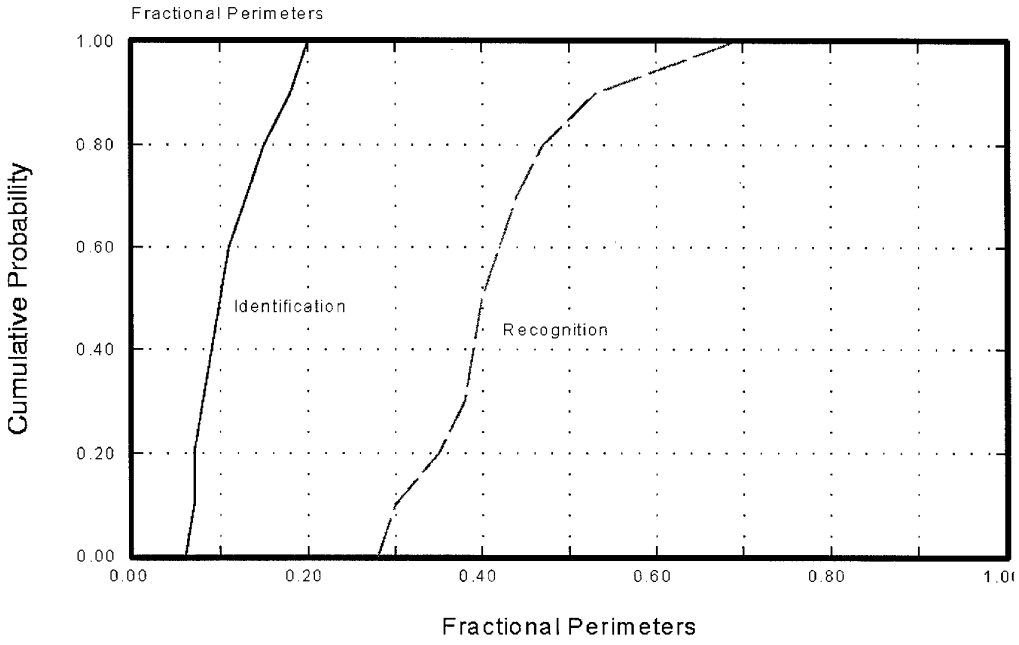


Figure 5. The distribution of fractional perimeter values required to model the threshold sizes for BAE recognition and identification experiments.

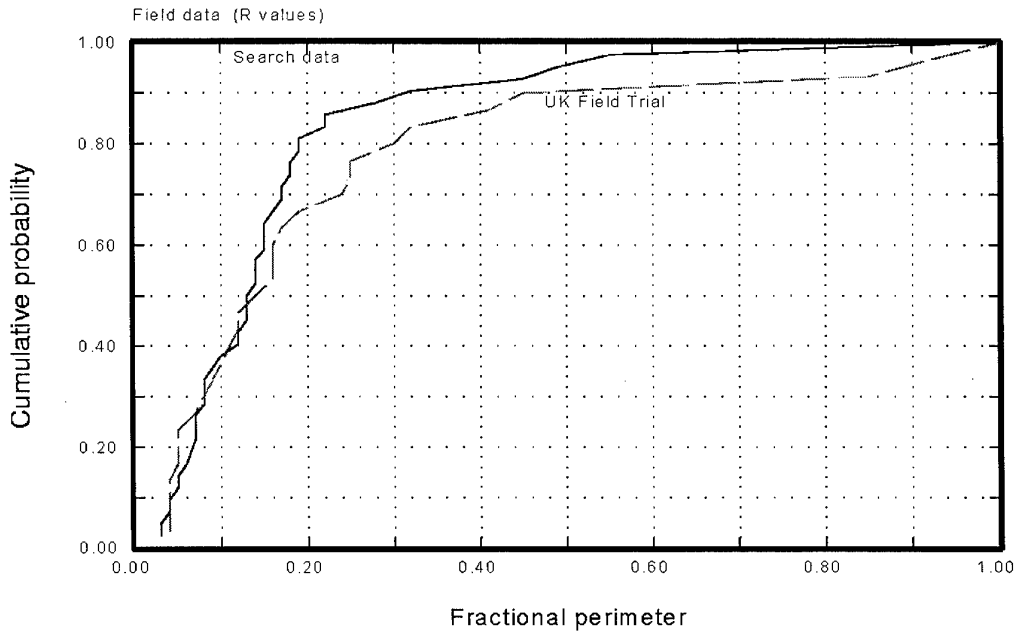


Figure 6. Fractional perimeter values for modelling SEARCH 2 and a UK field study.

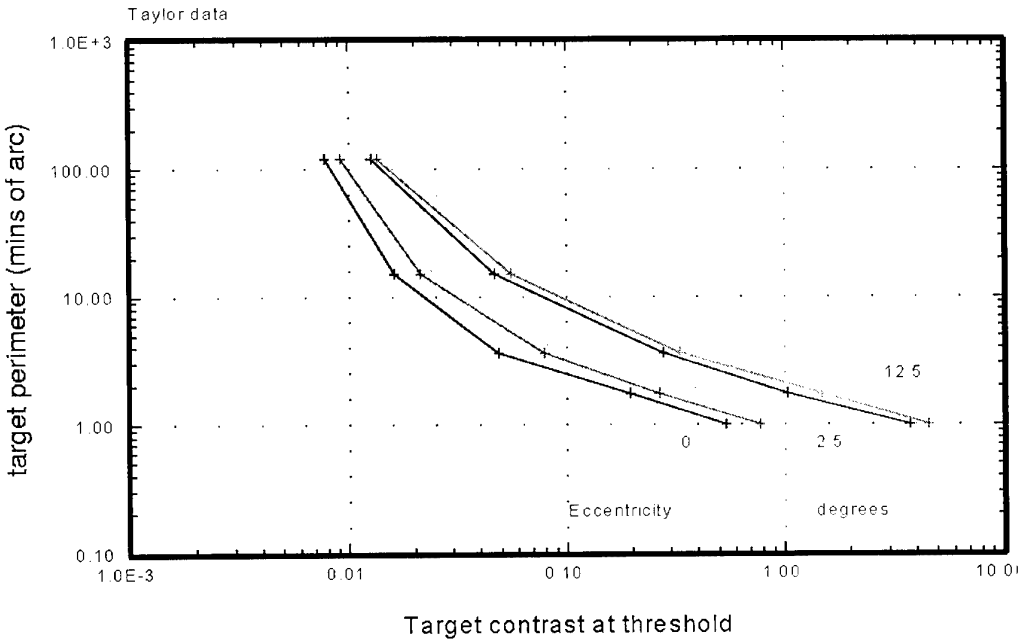


Figure 7. Peripheral contrast thresholds for varying size discs (Taylor , 1961)[10].

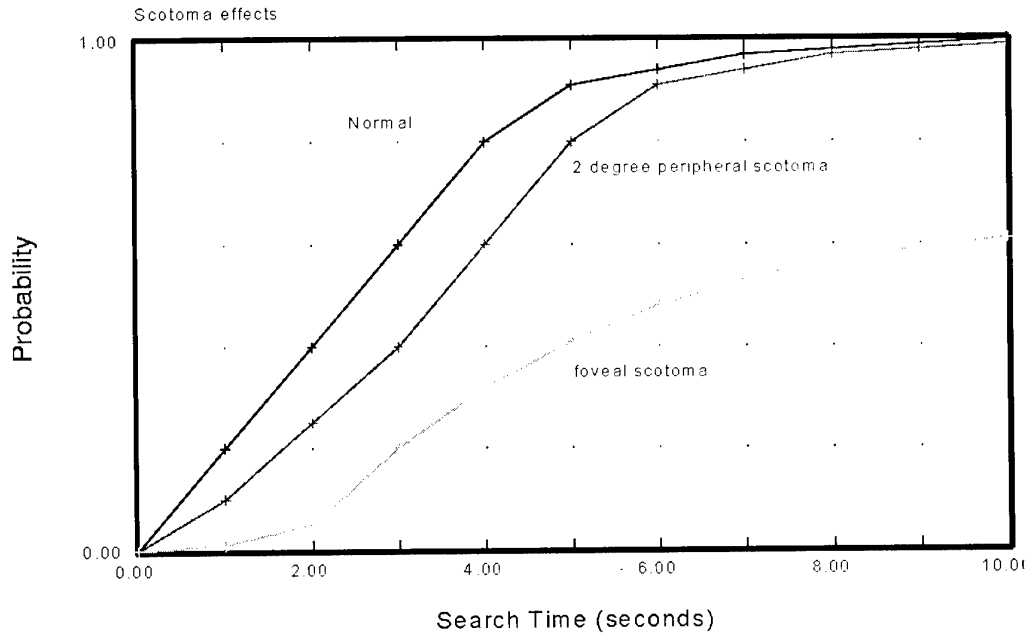


Figure 8. The effect of simulated 2 degree scotoma on search.

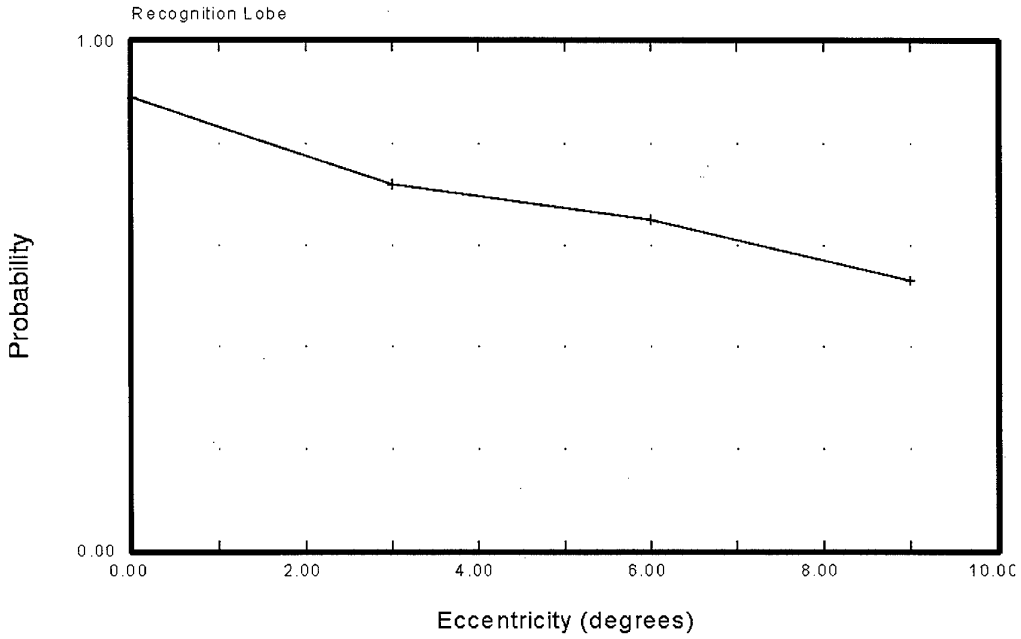


Figure 9. Peripheral recognition of targets in a structured scene for a single glimpse.

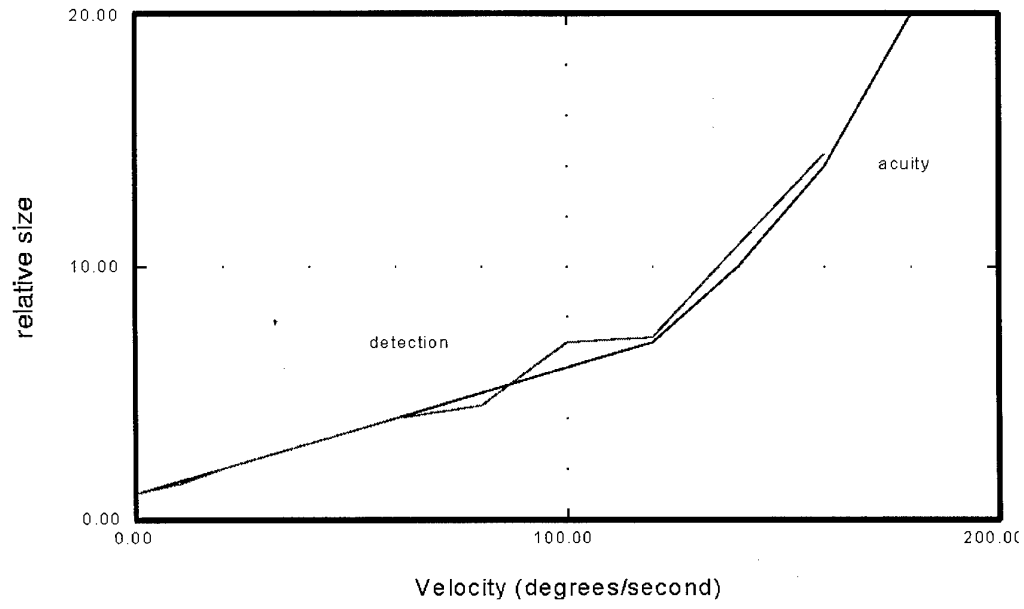


Figure 10. Motion Detection data from Barber and acuity with motion from Ludvigh and Miller Target Motion Data.

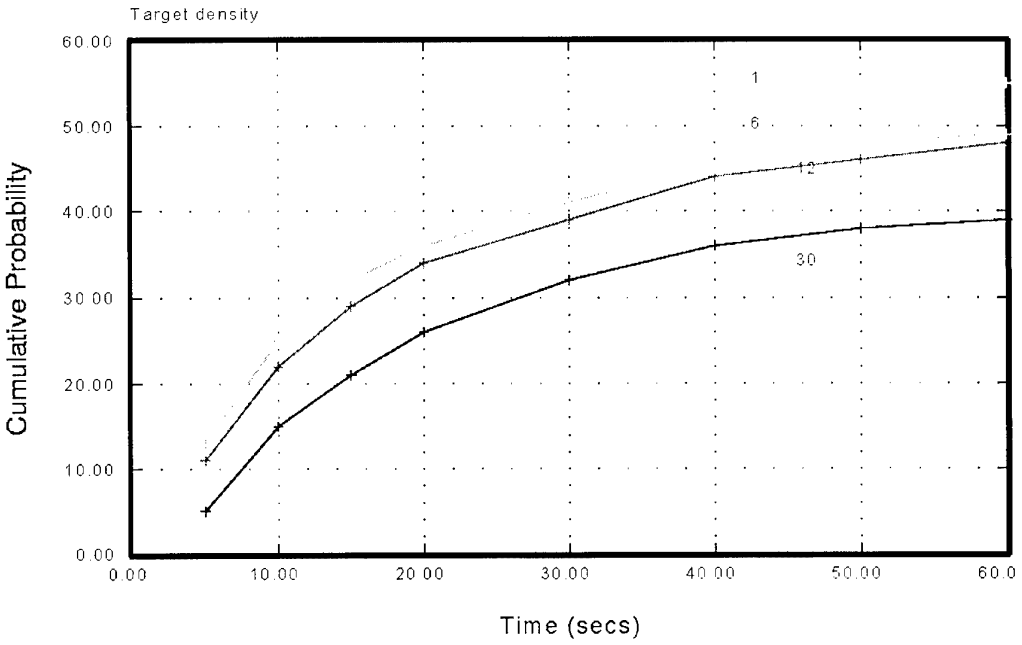


Figure 11. Search performance for varying target density (probability of finding all targets).

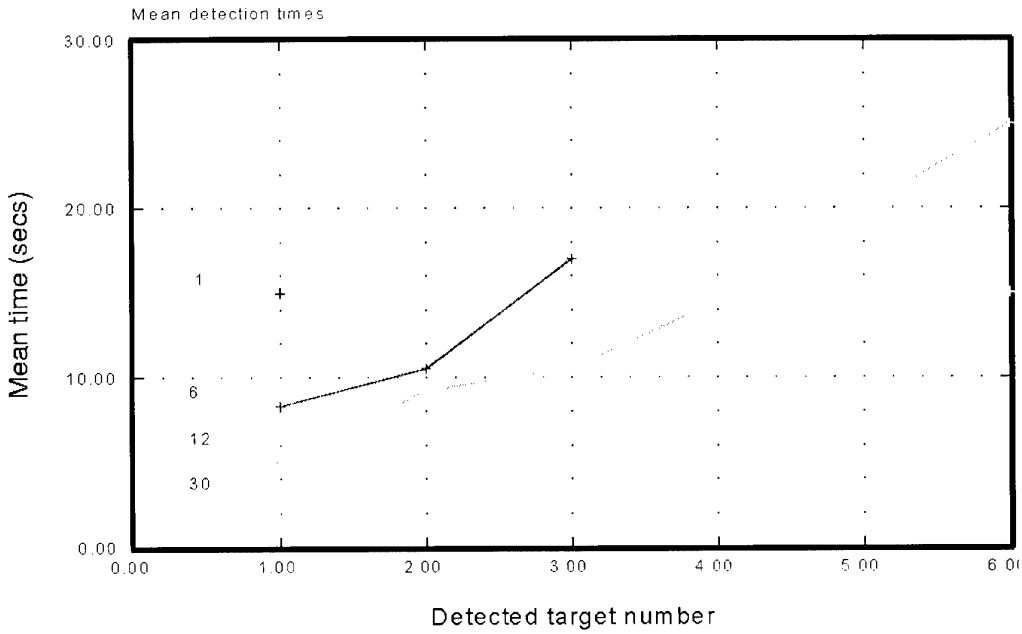


Figure 12. Search performance for varying target density.

IMAGE STRUCTURE MODELS OF TEXTURE AND CONTOUR VISIBILITY

Wilson S. Geisler, Thomas Thornton, Donald P. Gallogly & Jeffrey S. Perry
Department of Psychology and Center for Vision and Image Sciences
University of Texas at Austin, Austin TX 78712
geisler@psy.utexas.edu

1. SUMMARY

The perceptual mechanisms underlying texture and contour grouping/segregation play a dominant role in determining the visibility of targets in complex backgrounds. In most quantitative models of texture segregation the image is initially processed by channels selective along certain fundamental stimulus dimensions such as spatial frequency and orientation. These channels generally contain a nonlinearity, such as full-wave rectification, so that they signal the local contrast energy within the bandpass of the channel. Another stage of linear filtering, followed by a simple edge finding or thresholding mechanism, is then applied to the channel outputs to find the texture boundaries or regions. Although these channel-energy models have been successful in predicting texture segregation and discrimination performance for some classes of stimuli, there are large classes of stimuli that are readily segregated by human observers but which cannot be segregated by channel energy. The evidence suggests that more sophisticated models incorporating perceptual organization mechanisms will be required to predict human texture and contour segregation performance. This paper describes new experimental evidence, and a working model which, in principle, can account for a wider range of human segregation and grouping capabilities. The premise of the model is that the visual system typically extracts rich descriptions of local image structure, and that it uses these descriptions for subsequent segregation and grouping. The model contains physiologically-based low level mechanisms for extracting primitives, matching mechanisms for detecting structural similarity, and grouping mechanisms for binding structural parts into wholes. Quantitative predictions of the model for contour segregation performance are presented.

2. INTRODUCTION

"Bottom-up" mechanisms for grouping and segregation are absolutely essential to object detection and recognition. To recognize an object in a typical natural environment, the features of the object must be segregated, at least to some extent, from those of the surrounding objects and surfaces.

In recent years, most models of grouping/segregation have been based upon mechanisms which compare, across space, the contrast energy within spatial-frequency and orientation tuned channels (e.g., for reviews see, Bergen, 1991; Bovik, Clark, & Geisler, 1990; Graham, Beck and Sutter, 1992). While such grouping/segregation mechanisms may exist within the human visual system there are at least two grouping abilities they cannot explain. First, humans are able to segregate image regions based upon differences in the local spatial structure, even when the channel energies are the same (Thornton, et al. 1998; see later). Second, humans are able to detect spatial structure and statistical regularities that vary smoothly over space (i.e., non-stationary image structure). The most well-known example of this is the ability of humans to detect a contour formed by a sequence of line segments (a dashed contour) embedded in a background of randomly oriented line segments (e.g., Sha'ashua & Ullman, 1988; Field, Hayes & Hess, 1993; see later).

The aim of this paper is to demonstrate some of the weaknesses of the *channel energy models* and to demonstrate how some of those weaknesses might be addressed by models incorporating mechanisms which explicitly extract and use local image structure—*image structure models*. We begin by briefly describing a generalized channel energy model and an image structure model.

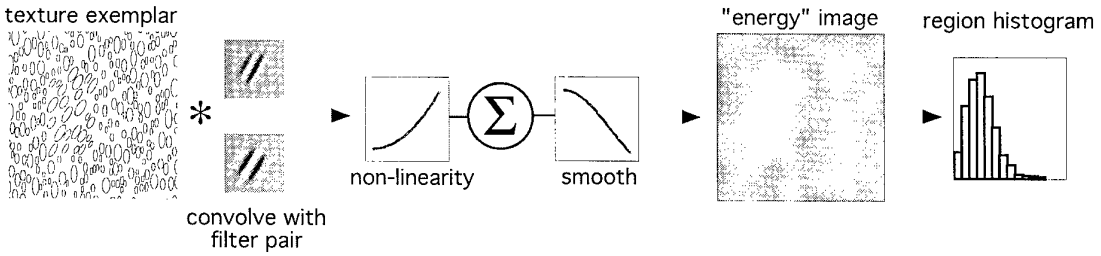


Figure 1. The generalized channel energy model for segregation and grouping. The input image is a complete set of filter pairs each tuned to a different spatial frequency and orientation. One filter pair is shown in the figure. The responses of each filter in the pair are squared and then summed, mimicking the response of a complex cortical cell in area V1. The responses (either with or without smoothing) form an "energy" image. In the most general case, the responses in a region are represented by a histogram of the response magnitudes taken over the region.

Next, we describe an experiment demonstrating that humans can easily segment large classes of textures which are impossible even for an optimal channel energy model. We also show that an image structure model can segment such textures. We then describe parametric measurements of contour detection performance and show that a very simple image structure model is able to account for most aspects of the data.

2.1 Generalized Channel Energy Model

Figure 1 illustrates the generalized channel energy model of segregation and grouping. In this illustration, the input image consists of a target region of diagonal ellipses in a background of vertical ellipses. The input is processed by 30 separate spatial-

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

frequency and orientation tuned channels (6 frequencies x 5 orientations), with spatial frequency bandwidths of 1 octave, and orientation bandwidths of 30 deg. The quadrature pair of receptive fields corresponding to one of the channels is illustrated in the second panel. The channel energy at each pixel location in the image is obtained by summing the square of the responses from the two quadrature components. One can think of the channel energy at a pixel as a response similar to the one that would be produced by a complex cell in primary visual cortex centered on that pixel. Next, the channel energy may be smoothed, or not smoothed, depending on the specific version of the model. As can be seen, the channel energy is greatest in the region of the image where the ellipse orientation is similar to that of the channel. In the generalized energy model, the responses in a region are represented by a response histogram tallied over all the pixel locations in the region. In this illustration, the response range was divided into 15 equal-width regions. In a more conventional energy model, the responses in a region are represented by the sum or average response in the region (i.e., the mean of the histogram). The generalized histogram model extracts some local spatial phase information and hence predicts that humans can discriminate a wider range of textures than predicted by a conventional energy model. Nonetheless, the experiment described later shows that there are many textures humans can segregate, which the ideal generalized channel energy model predicts cannot be segregated.

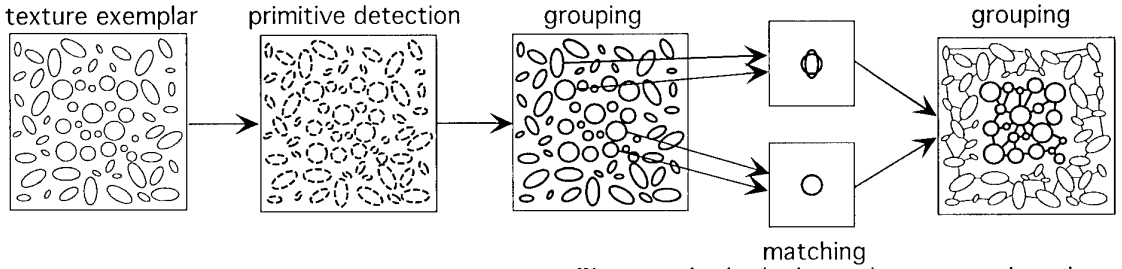


Figure 2. An image structure model for segregation and grouping. The input image is encoded into local primitives (in this case oriented line segments). Initial groups are formed by associative grouping. The initial groups are compared by a matching process to measure their similarities/differences. Higher order groups are obtained by another stage of simple or associative grouping. The processes of matching and grouping may be repeated.

2.2 An Image Structure Model

Figure 2 illustrates an image structure model of segregation and grouping. In this illustration, the input image consists of a target region of circles in a background of randomly oriented ellipses. The input image is encoded as a collection of local primitives (oriented pieces of contour). Different choices for the primitives are possible; in one version, we obtain the primitives by thresholding the responses of model simple cells tuned to different orientations (Geisler & Super, 1996/98). The primitives are represented by the small segments in the second panel. Next, associative grouping of the primitives (described below) is used to obtain initial groups, which, in this case, correspond to the "elements." The initial groups are then compared with one another by a matching process, under some family of transformations (e.g., translation, rotation and scaling). The matching process produces measures of the similarities/differences between the initial groups. In this case, the element shapes within the target and background regions match each other, but do not match across the regions. Finally, the grouping process is applied to the initial groups using the similarities/differences found by the matching process. Grouping on the basis of shape similarity will correctly segment the texture regions. This texture is difficult to segment on the basis of generalized channel energy because of the random sizes, positions and orientations of the elements. We now describe some of the more important components of the model in a bit more detail.

2.3.1 Similarity Measure

To determine the structural relationships between groups, there must be some mechanism for measuring the similarities (or equivalently, differences) of the groups along relevant stimulus dimensions. We assume this is done in parallel across the groups (e.g., across the "elements" in the third panel of Figure 2) by a matching process. In the current version of the model, we suppose that the matching process measures differences between the groups along stimulus dimensions which include form/shape (D_f), position (D_p), orientation (D_θ), size (D_s), symmetry (D_r), and continuation (D_c). The definitions of these *group differences* are described in more detail in Geisler & Super (1996/99). Later we give the definitions for position and continuation. Our working hypothesis is that the potential for binding the two groups together is given by the *total group difference*, which is a linear weighted sum of the group differences:

$$D = w_f D_f + w_p D_p + w_\theta D_\theta + w_s D_s + w_r D_r + w_c D_c \quad (1)$$

We suppose that the visual system has some control over the weights and hence may favor some dimensions under certain situations.

2.3.2 Simple and Associative Grouping

We assume that there are two basic grouping mechanisms which the visual system may use. The first is to bind together groups for which the total group differences are small. We call this *simple grouping*. More formally, if the total grouping difference (D_{ij}) between groups g_i and g_j falls below some criterion (β) then the groups are bound together:

$$if \quad D_{ij} < \beta \quad then \quad g_i \circ g_j \quad (2)$$

where the symbol "$\circ$" represents the operation of binding. In this definition, D_{ij} may represent the weighted sum of grouping differences for all the dimensions except position (proximity). For simple grouping we assume that the weight on the position difference is zero. This assumption is required in order for simple grouping to extend across significant spatial distances.

Associative grouping combines proximity grouping with simple grouping and a transitivity rule. Just as in simple grouping, if the total grouping difference between groups g_i and g_j falls below some criterion then the groups are bound together:

$$if \quad D_{ij} < \beta \quad then \quad g_i \circ g_j \quad (3)$$

In addition, if group i binds to group j and group j binds to group k then groups i and k are bound together:

$$if \quad g_i \circ g_j \;\&\; g_j \circ g_k \quad then \quad g_i \circ g_k \qquad (4)$$

In this definition, D_{ij} represents the total grouping difference, which includes the position/proximity grouping difference.

Our working hypothesis is that the visual system tries a number of values of the binding criterion either simultaneously or sequentially, and then picks values based upon three rules: the *stability rule*, the *performance rule*, and the *recognition rule*. The stability rule is a "bottom up" rule which depends upon the dynamics of group formation. When the binding criterion is varied there will be ranges in the value of the binding criterion where the pattern of grouping is changing rapidly and there will be ranges where the grouping is stable. The stability rule is to pick values from ranges where the grouping is stable (see Estabrook, 1966 for a similar concept of classification). The performance rule is a "top down" rule where the criterion is adjusted or selected based upon improving task performance. The recognition rule is a top down rule which depends upon feedback from subsequent recognition processes. The rule is to pick values of the binding criterion that yield recognized objects/parts when the groups are analyzed further; we suppose that this further recognition analysis is occurring simultaneously while the binding criterion is being varied.

2.3.3 Repeated Grouping

An important aspect of the image structure model is that the grouping processes are carried out repeatedly. First, associative or simple grouping is applied to the detected primitives to find initial groups. After matching, which provides new grouping differences, associative or simple grouping is applied again to obtain higher order groups. Although not indicated in Figure 2, there may be additional repetitions of matching and grouping. It is the repeated applications of matching and grouping which provide the detailed description of image structure.

3. EXPERIMENT 1

Thornton & Geisler (1998) showed that conventional channel energy models predict that certain classes of texture are impossible to segregate when, in fact, humans find them easy to segregate. The classes of textures they considered are similar to those in Figure 2 (see also Victor & Brodie, 1978). The target and background regions consisted of elements that differed in shape, but were randomized in orientation, size, and position.

In the present experiment we tested whether similar results hold for generalized channel energy models, which assume that the human visual system uses the additional information which is contained in channel response histograms computed over texture regions. The generalized channel energy models are more powerful and hence more difficult to reject. To test these models we developed a special procedure for constructing the texture segregation stimuli.

3.1 Methods

The logic of the experiment is quite simple. Construct texture segregation stimuli for which the optimal generalized channel energy model is at chance performance in a forced choice task. If the human observers can perform the segregation task at above chance then the general class of channel energy models is rejected. The difficult part is in constructing the stimuli.

3.1.1 Task

The task was to decide whether a rectangular target region, filled with one texture, was oriented vertically or horizontally within a background region, filled with another texture. The location of the target region was random from trial to trial.

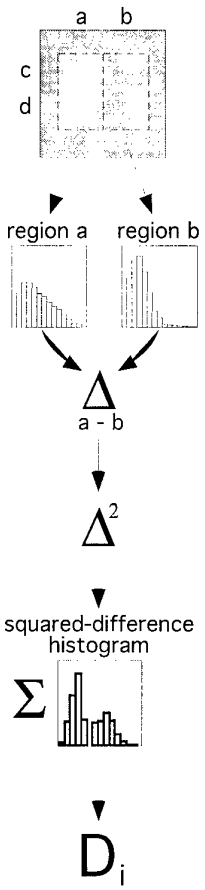


Figure 3. Initial processing steps of the general channel energy model applied to the segregation task of Experiment 1. For each channel-energy "image" a histogram is computed for regions a and b. The histograms are differenced, squared and summed to obtain a measure, D_i, of the histogram difference for each channel.

3.1.2 Stimuli

The exemplars for any given texture stimulus were created by filling a rectangular "target" region (2 by 4 deg) with elements of one shape, and filling the square "background" region (8 by 8 deg) with elements of a another shape. The shapes were always smoothly connected contours. They were generated by summing sinewave components with frequencies in orientation that were harmonics of one cycle per 360 deg. Different random shapes were obtained by randomly selecting the radial amplitudes and phases of the components and then filtering (i.e., multiplying the amplitudes by a transfer function). Exemplar texture stimuli were created by filling each region (target or background) with non-overlapping elements having a single shape, but random size, position, and orientation. For the purposes of creating the stimuli, each element was represented by a virtual circumscribing box. The virtual boxes were randomly placed one at a time within the image, with the restriction that if a new virtual box overlapped an

existing one then a new element was selected. The center of the circumscribing box was allowed to just touch the invisible border defining the regions; this resulted in texture regions with a "natural," irregular appearing boundary. In order to insure that segmentation of target and surround was due solely to specified regional differences in local shape, first order cues of luminance and contrast were balanced across regions and exemplars by keeping pixel density constant.

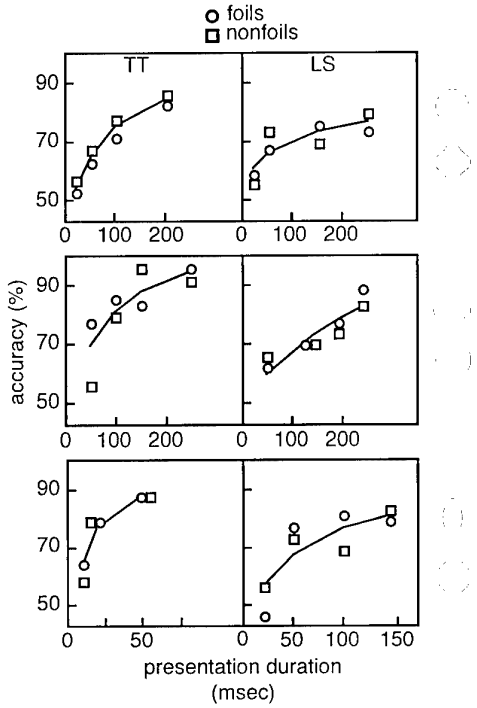


Figure 4. Texture segregation accuracy as a function of stimulus duration for two subjects. The shapes on the right indicate the shapes of elements in the target and background regions. The size, orientation and position of the elements in each region were random. The optimal channel energy model is at chance performance for these stimuli.

From many texture exemplars generated in the above fashion, we selected a subset for which the generalized channel energy model was at chance. To do this we simply collected exemplars that were misclassified (i.e. "foils"). These "foils" were then combined with a number of correctly-classified exemplars to form a stimulus ensemble that by definition yields 50% correct performance. In the experiments reported here, all stimulus ensembles were selected in this manner.

The key step in generating the stimulus ensembles was determining the optimal performance for the general channel energy model in the segregation task. Figure 3 illustrates the initial sequence of processing for the segregation task. On each trial, a set of 30 energy "images" were computed as per the process described in Figure 1 (one image for each of the channels comprising the model's front-end). The first panel in Figure 3 represents one of these energy images. To be conservative, we determined model performance assuming that the target could appear in only one of four locations: *a*, *b*, *c*, and *d* in Figure 3. (The human observers were confronted with greater uncertainty because the targets were randomly positioned within the central region.)

Optimal performance was achieved by applying a maximum likelihood decision rule to the channel responses produced on each trial. First, we computed the sum of the squared difference in the histograms for regions *a* and *b*, for each channel. In Figure 3, D_i represents the value of this quantity for the i^{th} channel. Then, we computed the probability that these 30 values were generated by a vertically oriented target region and by a horizontally oriented target region. The response was which ever orientation was more probable.

The probability densities for the two orientations were assumed to be adequately represented by 30-dimensional multivariate normal density functions (one dimension for each channel) with arbitrary mean vectors and covariance matrices. The mean vector and covariance matrix for vertical targets was estimated from a large number of exemplar vertical stimuli. Similarly, the mean vector and covariance matrix for horizontal targets was estimated from a large number of exemplar horizontal stimuli.

3.1.3 Procedure

Observers made forced choice target orientation decisions ("vertical"/ "horizontal") to individual texture exemplars presented in blocks of 96 trials. Target orientation was random and balanced across blocks. All texture stimuli were presented briefly at maximum contrast, and were followed by a matched pattern mask and feedback tone. Presentation duration was varied across blocks to obtain psychometric functions.

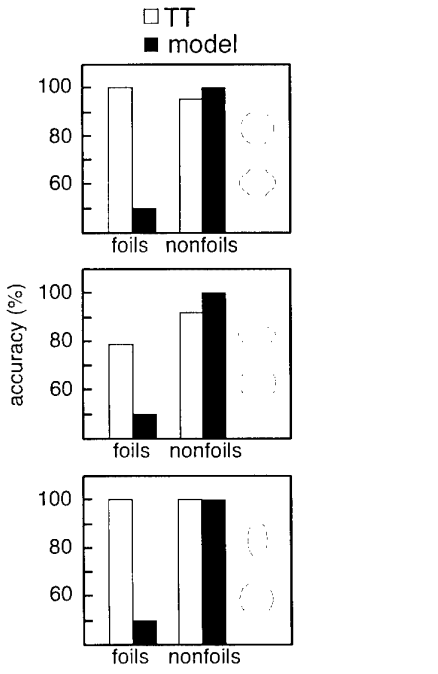


Figure 5. Segregation accuracy for three texture stimuli, for a human observer and an optimal generalized channel energy model. Stimulus duration was 200 ms.

3.2 Results

Figure 4 shows the texture segregation performance for two subjects on three different textures. The shapes of the elements in the textures are shown on the right. The results demonstrate that even for brief presentations the human observers are performing better on these stimuli than the optimal channel energy model.

However, it is possible that during the course of the experiment subjects are learning and using histogram differences for the specific stimuli in the experiment. To test this possibility, we created stimulus ensembles that consisted of approximately 10% foils and 90% non-foils. We computed the optimal histogram model for these particular stimuli and then determined the performance of the optimal model on these same stimuli (which overestimates of the model's performance). Figure 5 shows the performance of the model and one of the subjects on these ensembles. The subject outperformed the optimal channel histogram model, even though the model was being given an unfair advantage.

3.3 Discussion

This experiment demonstrates that human observers can, in brief presentations, segregate image regions based solely on differences in local image structure. This result is undoubtedly quite general because the texture elements in this experiment were picked arbitrarily. Indeed, we have similar preliminary data for other element shapes.

Although generalized channel energy models represent some information about local image structure (i.e., phase information), they do not represent sufficient structural information to segregate the class of stimuli described here. This creates difficulties for all channel energy models proposed to date, because these models do not segregate as accurately as the optimal generalized channel energy model considered here.

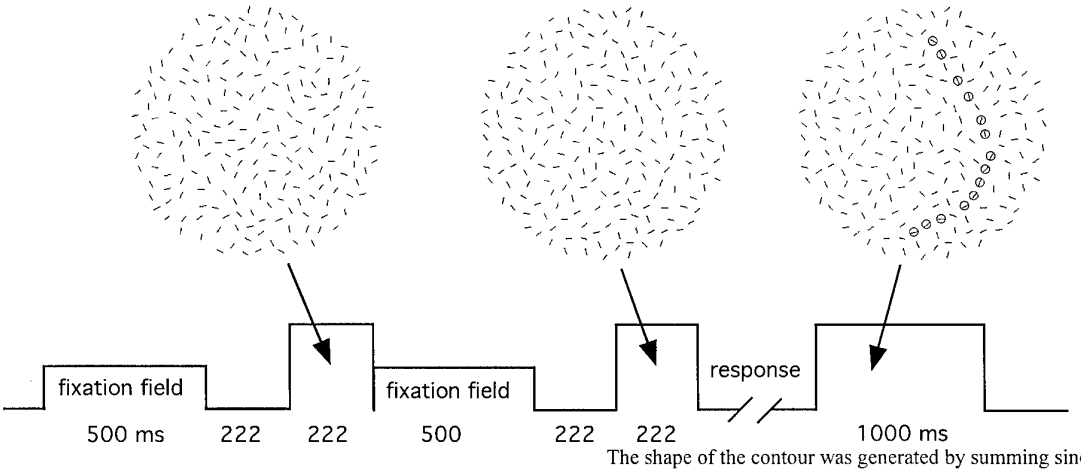


Figure 6. Example stimuli and presentation sequence for the contour detection experiment (Experiment 2).

Of course, it is intuitively obvious that the human visual system extracts precise descriptions of local spatial structure. What the experiments here demonstrate is that the visual system uses such descriptions to perform fast ("preattentive") region segregation. Although not demonstrated here, image structure models can segregate the kinds of textures considered in this experiment.

4. EXPERIMENT 2

The second major class of tasks that pose a difficulty for channel energy models are those that involve grouping of regions containing smoothly changing image structure, such as smooth contours. To obtain additional systematic data on human ability to group contour information, and to provide a test of the image structure model, we measured contour detection performance as a function of contour shape and length. This is a particularly useful task because it involves complex naturalistic judgements under high degrees of uncertainty; yet, the predictions depend upon very few parameters in the image structure model.

4.1 Methods

Accuracy was measured in a two interval forced choice task for detection of line-segment contours in a background consisting of randomly oriented line segments. Four properties of the randomly shaped contours were parametrically varied: amplitude, fractal exponent, length, and level of orientation jitter of the contour elements. This family of contours was selected to be representative of a broad range of naturalistic contours.

4.1.1 Stimuli

Figure 6 illustrates the time line for a single trial, including examples of a background and a background + target. The circular display was 12.3 deg in diameter (32.5 pixels/deg), at the viewing distance of 112 cm. The line-segment elements were 0.31 degrees in length. The target contour and background texture were created in a fashion similar to that in Experiment 1. For purposes of creating the stimuli, each line element was represented by a virtual circle with a diameter equal to twice the element length. The virtual circles were randomly placed one at a time in the image with a restriction that if a new virtual circle intersected an existing one then a new random position was selected.

The shape of the contour was generated by summing sinewave components with frequencies that were harmonics of 0.5 cycles per image. The sinewave components always modulated about an axis through the center of the display; the orientation of the axis was random on each trial. Different random contour shapes were obtained for each trial by randomly selecting the amplitudes and phases of the components, and then filtering (i.e., multiplying the amplitudes by a transfer function). The line elements were randomly placed on the contour first. Then the background line elements were added such that the density of line elements in the background was the same as along contour.

Contour detection accuracy was measured parametrically as a function of four variables: (a) the fractal exponent of the amplitude transfer function (1, 1.5, 2, and 3), (b) the RMS amplitude of the contour modulation (6.5%, 12.5%, 25%, and 50% of the display diameter), (c) the contour axis length (20%, 40%, 60% and 80% of the display diameter), and the range of orientation jitter of the elements (0, 30%, 50%, and 70% of the maximum value, 180°).

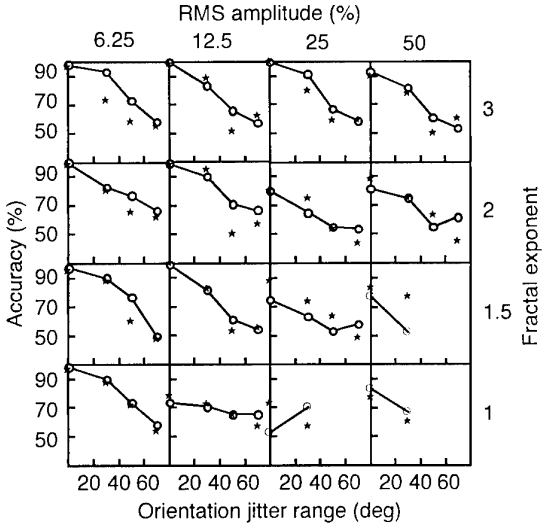


Figure 7. Length = 80% of display diameter. Open circles: contour detection accuracy for random contours, as a function of contour shape (fractal exponent), average contour amplitude (RMS amplitude), and magnitude of orientation jitter of the elements (Orientation jitter range). Solid stars: predicted performance of a two-parameter image structure model.

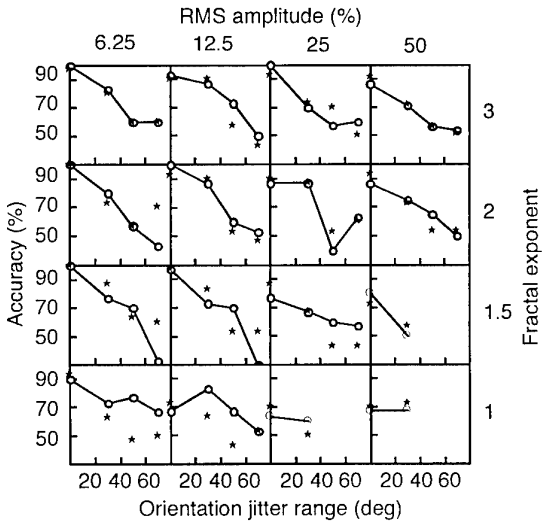


Figure 8. Length = 60% of display diameter. See Figure 7.

4.1.2 Procedure

As shown in Figure 6, on each trial the fixation cross was extinguished 222 ms before presentation of the two test intervals, which were each 222 ms in duration and separated by 720 ms. After the subject responded, he was informed about the correctness of the response, and shown the actual location of the contour.

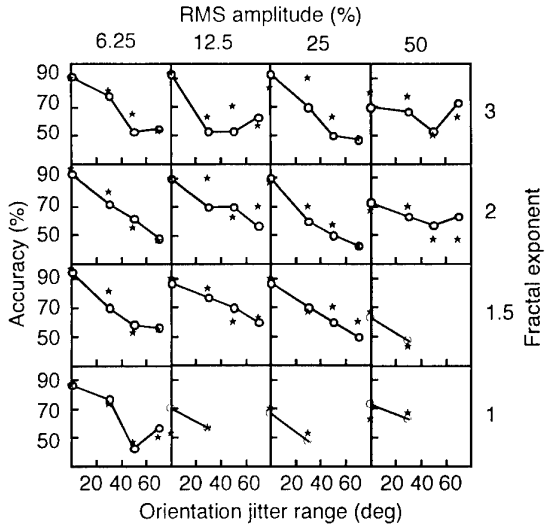


Figure 9. Length = 40% of display diameter. See Figure 7.

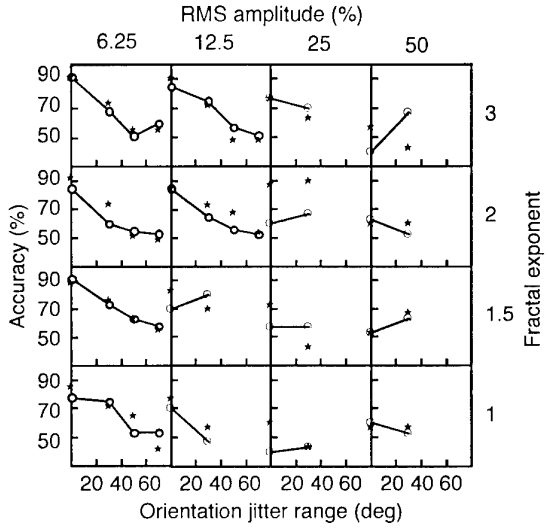


Figure 10. Length = 20% of display diameter. See Figure 7.

Each experimental session consisted of 16 blocks of 30 trials and lasted approximately 40 min. In each block, the stimulus parameters were held fixed. The order of conditions was picked to minimize systematic practice and fatigue effects. All conditions were repeated twice for a total of 60 trials per condition per subject.

4.2 Results

Figures 7-10 show the results for two subjects. Each figure is for a different contour length. The open circles in each panel within a figure show the average accuracy for the two subjects as a function of the range of element orientation jitter. (At the time of writing, the data for this experiment were not complete, so some data points represent results for only one subject.) The solid stars show the predictions of an image structure model described below. Across rows, the plots are for contours of different fractal exponents. Across columns the plots are for contours of different RMS amplitudes. The panels with only two points indicate conditions where all non-zero levels of jitter were not tested, based

upon pilot experiments showing that performance was poor even with 0% jitter.

There are some obvious trends in the data. Performance generally improves with increases in the fractal exponent and contour length, and generally declines with increases in RMS amplitude and jitter. Although these trends are not surprising, the specific levels of performance in the different conditions provide strong constraints on models of contour detection.

4.3 Discussion

This experiment provides a parametric overview of human capabilities for detecting contours in noisy backgrounds. As the data show, humans are quite good at detecting contours even when there is great uncertainty in location, orientation, and shape of the target contour. For example, contours like that in the middle picture of Figure 6 are detected with better than 90% accuracy.

It is generally acknowledged both by perception and computer vision researchers that humans have a remarkable ability to detect spatial structure and statistical regularities in images, even when the structure is unfamiliar (e.g., see Witkin and Tannenbaum, 1982). Thus, our initial assumption was that the relatively simple class of image-structure models described here would be unable to achieve the performance levels of humans in the present experiment. Our aim in testing the models was to identify conditions where the models fail to predict performance, with the hope that these failures might provide hints about what additional mechanisms the human visual system may be using. To our surprise, a simple image structure model (with just two free parameters) does a good job of accounting for the data, suggesting that human contour detection performance in these types of displays may be largely explained by relatively simple, essentially "bottom-up," processing.

In the specific model described here, the input was taken to be all the individual line segments in the display (not the individual pixels). This is equivalent to assuming that primitive detection and initial grouping have already found the groups corresponding to the individual line segments. Thus, the primary computations in the model consisted of a matching stage which compared line segments and a subsequent grouping stage which bound them into groups (see Figure 2).

Because the line segments were all of the same size and form, equation (1) reduces to the weighted sum of the position difference, orientation difference and continuation difference. Furthermore, we found that the information extracted by continuation difference (which had not been considered in Geisler & Super, 1996/99) was redundant with the orientation difference, so we could eliminate the orientation difference. Thus, equation (1) reduces to:

$$D = w_p D_p + w_c D_c \qquad (5)$$

The value of D was computed for each possible pairing of line segments in the image. Groups were then formed by applying associative grouping for a particular value of the binding criterion, β. In other words, we computed D_{ij} for each possible pairing of line-segment groups, g_i and g_j, and then applied the associative grouping rules given by equations (3) and (4). This was done for the stimuli in both intervals in the forced choice presentation. The longest group obtained in each interval was then selected. Which ever of these two groups had more elements was picked as the interval containing the contour. If the two groups happened to have the same number of elements then the group with the smallest summed grouping differences (the strongest binding) was picked as the interval containing the contour.

There are only two parameters in this model: one of the dimensions weights, w_p, and the binding criterion, β. The other dimension difference weight, w_c, is not free because the weights sum to 1.0.

We now define the group difference measures. These particular measures were devised on the basis of intuition, and a little trial and error. No great significance should be attached to these specific formulas. Undoubtedly, there are other related formulas which would capture the same information about the differences between line segments.

The position difference, D_p, was taken to be the Euclidean distance between the nearest pixels in the two line segments. We used this measure based upon experimental results of Geisler & Super (1996/99), who found that proximity grouping is better described by the near point distance than by the distance between object centroids. Note that because of the position difference component, the computation of the total difference, D, is a local process; only neighboring line segments influence the groups that are formed.

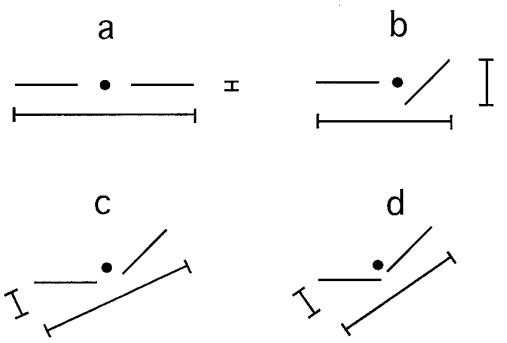


Figure 11. Illustration of the continuation difference measure. Each sub-figure shows a pair of line segments in some configuration. The solid dot shows the centroid of the group formed by the two segments. The long "error bar" shows the orientation of the major axis; the short "error bar" the orientation of the minor axis. The position difference is the ratio of the standard deviations computed along the two axes (through the centroid); this quantity is approximately the length of the short "error bar" divided by the length of the long "error bar."

The continuation difference, D_c, is a straight-forward measure meant to capture (in conjunction with the position difference measure) the degree to which two line segments could be smooth continuations of each other. Note first that any two line segments can be considered as a single group of points. This group will have a centroid, and a major axis, which is the best fitting line passing through the centroid. The continuation difference is defined to be the ratio of the standard deviation of the distance of the points from the major axis divided by the standard deviation of the distance of the points from the minor axis (the axis perpendicular to the major axis):

$$D_c = \frac{\sigma_M}{\sigma_m} \qquad (6)$$

Note that this measure varies between 0 and 1. Figure 11 illustrates how this measure behaves. The longer "error bar" in each sub-figure indicates the orientation of the major axis, the shorter "error bar" the orientation of the minor axis, and the solid dot the centroid. The continuation difference is roughly proportional to the ratio of the length of the short "error bar" to the long "error bar." When the two line segments fall along a straight line (Figure 11a) then the standard deviation about the major axis is zero, and hence the continuation difference is zero. If the second line segment is rotated (Figure 11b) the standard deviation about the major axis increases and so does the continuation difference. If the second line keeps the same orientation difference but is shifted vertically so that it is more consistent with a smooth contour (Figure 11c), then the continuation difference decreases. For a given orientation difference between the line segments, the closer the line segments the greater the continuation difference (compare Figures 11c & 11d). This is consistent with the fact that a greater curvature would be required to connect the two line segments when they are closer.

The model contained one additional constraint. Although the contours were random in orientation, shape, and position on every trial in a block, there was always some limitation on the possible locations of the contours. These limitations were obvious to the subjects when running the experiment. The model was also given this information; it did not consider groups which fell outside the region of possible contour locations.

To estimate the best fitting parameter values, a coarse grid search was followed by a more refined grid search at the most promising locations. For each pair of parameter values, the performance of the model was computed for exactly the same stimuli that the subjects saw.

The solid stars in Figures 7-10 show the predictions of the model. As can be seen, the model does a remarkably good job of predicting the performance across all the conditions. The estimated parameter values are as follows: $w_p = 0.2$, $w_c = 0.8$, $\beta = 0.17$.

Importantly, these parameter values, which fit the human data best, are also the values that maximize the absolute accuracy of the model in the task. In other words, combining local measurements of distance and continuation so that they produce the greatest accuracy in the model, also yields a model performance that is close to human performance. This fact adds some support for this class of models.

Further, the results suggest that human contour detection in these types of displays may involve only local measures of group differences followed by an unsophisticated grouping mechanism, such as associative grouping. Something similar to the local grouping difference measures and associative grouping should be relatively easy to implement neurally.

Finally, if the human visual system is using these simple mechanisms then the results imply that it has evolved or learned nearly optimal weights for combining local difference information and nearly optimal criteria for controlling associative grouping.

5. CONCLUSION

The experiments and analyses reported here demonstrate that there are many texture grouping and segregation situations that are difficult to model within the framework of the generalized channel energy models. The heart of the difficulty for such models is that the human visual system extracts detailed local spatial structure and is able to use it for grouping and segregation. Unfortunately, we see no way to avoid the difficult problem of modeling how the visual system extracts and represents spatial structure. Although image structure models cannot yet be easily applied to arbitrary images, we have made a start, and we have demonstrated that they can be quite effective in limited domains. In particular, a very simple image structure model that combines local measures of proximity and continuation can account for human ability to detect random contours that are representative, in complexity and uncertainty, of those occurring in the natural environment.

6. REFERENCES

Bergen, J. R. "Theories of visual texture perception." In D. Regan (Ed.), *Vision and visual dysfunction* (Vol. 10B: Spatial vision, pp. 114-134). New York: Macmillan, 1991.

Bovik, A. C., Clark, M., and Geisler, W. S. "Multichannel texture analysis using localized spatial filters." *Pattern Analysis and Machine Intelligence, 12*, 55-73, 1990.

Estabrook, G.F. "A mathematical model in graph theory for biological classification." *Journal of Theoretical Biology*, 12, 297-310, 1966.

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vision Research, 33*(2), 173-193.

Geisler, W.S. and Super, B.J. Perceptual organization of two-dimensional patterns. *Psychological Review*, under review.

Geisler, W.S. and Super, B.J. Perceptual organization of two-dimensional patterns. UT-CVIS-TR-96-002, Austin, Texas: Center for Vision and Image Sciences, 1996.

Graham, N., Beck, J., and Sutter, A. "Nonlinear processes in spatial-frequency channel models of perceived texture segregation: effects of sign and amount of contrast." *Vision Research, 32*(4), 719-743, 1992.

Sha'ashua, S., and Ullman, S. *Structural saliency: The detection of globally salient structures using a locally connected network.* Paper presented at the Proceedings of the Second International Conference on Computer Vision, 1988.

Thornton, T. and Geisler, W.S. "Texture segregation on the basis of local shape information." *Investigative Ophthalmology & Visual Science Supplement*, (ARVO) 39/4, S649, 1998.

Victor, J.D. and Brodie, S.E. "Discriminable textures with identical Buffon needle statistics." Biological Cybernetics, 31, 231-234, 1978.

Witkin, A. P., & Tenenbaum, J. M. On the role of structure in vision. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and Machine Vision* (pp. 481-543). New York: Academic Press, 1983.

7. ACKNOWLEDGEMENTS

This work was supported by grants from the National Eye Institute, NIH.

COMPARING HUMAN TARGET DETECTION WITH MULTIDIMENSIONAL MATCHED FILTERING METHODS

[1]Krebs, W.K., [2]Scribner, D.A., [1]McCarley, J.S., [1]Ogawa, J.S., and [1]Sinai, M.J.
[1]Naval Postgraduate School, Department of Operations Research, 1411 Cunningham Road, Monterey, CA 93943
[2]Naval Research Laboratory, Optical Sciences Division, Washington, D.C.
E-mail: wkrebs@nps.navy.mil

1. SUMMARY

Recent technological advances in sensor manufacturing enable the use of separate spectral bands; e.g., MWIR and LWIR, to generate spatially registered imagery. Human factors experiments can be used to test whether a sensor can improve operator performance for detecting or recognizing a target[1]. Although human factors experiments are of tremendous value, these tests are time consuming and resource intensive. In order to reduce costs associated with collecting behavioral data, an alternative approach is discussed. We propose using signal detection theory, to compliment and reduce the amount of classical human performance testing. As a test case we have studied whether multi-spectral sensors are significantly better than single band sensors.

Scribner, Satyshur, and Kruer (1993) demonstrated that a two-dimensional matched filter (spatial) optimized for a specific target and background power spectra, can be used to estimate an observer's ability to detect the target embedded in a cluttered background. Three different background images were used with, and without, a target present. False alarm and target detection probabilities were computed and results were plotted on a Receiver Operating Characteristic (ROC) curve. The matched filter ROC curves were then compared to behavioral ROC curves. Results showed that the matched filter ROC curves were similar to behavioral ROC curves with color fusion and long-wave infrared showing the highest sensitivity and mid-wave and short-wave infrared scenes were significantly less sensitive. These results indicate that the matched filter analysis may be used to model human behavior.

Keywords: Signal Detection Theory, Matched Filter Analysis, Receiver Operating Characteristic, Human Performance Modeling, Target Detection

2. INTRODUCTION

Military applications require the use of various sensors to determine operational threats and opportunities. The combination of such sensors promise to provide an account of the opposition that is superior to those of individual sensors that operate at particular wavebands. It is desirable to choose the optimal types and combinations of sensor information that are maximally responsive to target types likely to be encountered in the field. This assessment must be done under realistic physical and psychophysical circumstances. Furthermore, it is desirable that the information obtained be modeled productively, i.e. so that experimental results can be interpolated and extrapolated near the conditions under which they are obtained.

This study will modify an existing matched filter model[2] to fit meaningful human performance metrics that can be revised and extended where necessary to represent the data obtained during field tests. This model will be used to evaluate fused imagery systems requirements and performance. Furthermore, this model will indicate what type of data will be needed to validate the type of sensor fusion data to be collected in the future.

In order to assess an operator's ability to detect a target while viewing sensor imagery, different disciplines have developed methodologies to measure operator performance. The Night Vision and Electronic Sensors Directorate (NVESD) has developed analytical models to predict target detection ranges for sensors that operate in the visible and infrared bands[3,4]. These electro-optical models provide an adequate prediction of a user's ability to detect a target at any given range. In order to improve the validity of the models, atmospheric conditions, sensor characteristics, target characteristics, clutter, estimated time that an operator searches for the target, and an assortment of other parameters are used to model human performance. Currently, these models are limited to single-band sensors; however, the next generation models may incorporate multi-spectral sensor performance.

Recent technological advances in the design and manufacturing of multi-spectral sensors now allows spatially registered imagery to be mapped to a high speed processor where it can be fused and displayed to an end user[5]. Within the last several years, numerous groups have developed sensor fusion algorithms[2,6-10] that may improve operator performance. These techniques may differ on the algorithm approach, but they all have the same objective: improving the image quality for the observer. Several behavioral studies[1, 11-16] and image quality studies[2,10] have tried to quantify the benefits of sensor fusion, but the results were inconsistent. This is not surprising considering that in many cases different spectral bands were used and a number of other parameters varied as well, such as camera sensitivity, and target and background characteristics.

Tanner and Swets (1954) proposed that statistical decision theory may be used to predict operators decision behavior. Signal Detection Theory is a common technique used by vision scientist to measure subjects' sensitivity and response bias to a set of stimuli[18]. Whether target detection is accomplished through the human visual system or by means of a matched filter, the theory of signal detection requires recognizing a signal plus noise from a steady state noise background. Vision scientists use signal detection theory to measure operator performance to an assortment of stimuli. Similarly, an image-processing algorithm may use a matched filter technique that is based on signal detection theory to predict operators' performance through a sensor. Ideally, the Receiver Operating Characteristic (ROC) plots derived from both methodologies should yield similar results. The advantage of the matched filter technique allows the system engineer to conduct multiple simulations for a wide variety of backgrounds and target types. These simulations require minimal resources compared to costly human performance field tests.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

A matched filter is a two-dimensional (2-D) array, which has been optimized to maximize the signal-to-noise ratio and provide a measure of the spatial correlation between the input image and the reference image[19]. The resulting filters are "tuned" to negate the effects of the background clutter and other noise sources in the image. A matched filter is the optimum linear filter for the detection of the target. Scribner et al. (1993) used a matched filter on long-wave infrared (9.0 to 11.6 μm), medium-wave infrared (4.5 to 5.5 μm), and short-wave infrared (2.0 to 2.6 μm) sensors, as well as on a fused single image of these bands. In this approach, spatial-only and spatial-spectral matched filters were derived for the three infrared images and the fused composite image respectively. The intent of these matched filters was to simulate the detection ability and sensitivity of the human visual system. Although the matched filter is commonly used within the physics and engineering communities to quantify sensor performance, it has been used to some extent by the medical field to detect tumors in a x-ray image[20].

The objective of this paper is to compare and contrast behavioral and matched filter ROC plots to determine whether the matched filter technique is a good predictor of human performance. The advantages of the matched filter model are threefold. First, it provides a sensor image fusion metric that can be used to evaluate different sensors. Second, it quantifies the degree of "enhancement" achieved by a fusion process, thus allowing for direct comparisons of the various sensor fusion algorithms. Third, it may have the ability to predict human visual performance across a variety of background and target conditions.

A standard visual search paradigm will be conducted for three different multi-spectral natural scenes. In experiment 1, behavioral ROC plots will be compared to the matched filter ROC plots to determine whether the matched filter technique accurately predicts observers' sensitivity. In experiment 2, eye movement data will be recorded to determine whether observers' scan pattern correlates with the ROC analysis. It is hypothesized that observers viewing a low contrast stimulus will exhibit longer saccade lengths and shorter fixations as well as show a low sensitivity for detecting the target (i.e., less correct responses and more errors).

3. EXPERIMENT 1

3.1. Behavioral Test

Subjects: Fourteen male military officers (mean age 31.7 years old) participated in this visual search study. All subjects had normal (20/20), or corrected to normal, acuity and color vision. Subjects were naive to the purpose of the experiment and none had participated in previous visual search experiments. All subjects signed an informed consent and were briefed on the ethical conduct for subject participation in the Protection of Human Subjects[21].

Apparatus: Stimuli were presented by a VisionWorks computer graphics system[22] on an IDEK MF-8521 high-resolution color monitor (21" X 20" of viewable area, .28mm dot pitch) equipped with a non-glare, anti-reflect, P-22 phosphor. The monitor's resolution was 800 by 600 pixels (x=75.02 and y=74.92 pixels/degree), 98.9 Hz frame-rate, mean chromatically of Y= 50.2, x = 0.334, y = 0.336 (1931 CIE), and a maximum luminance of 100 cd/m^2. Luminance of the monitor was linearized by means of an 8-bit look-up table for each of the red, green, and blue guns. Subjects viewed the monitor from 1.5 meters and were positioned by an adjustable chinrest. Subjects viewed the stimuli under mesopic conditions.

Stimuli: Three single-band stimuli (short-, mid-, and long-wave infrared) and two composite stimuli (fused-color and fused-gray) were selected from a multispectral natural scene database. The selection criteria consisted of scenes that contained heterogeneous terrain characteristics and no man-made targets (figure 1).

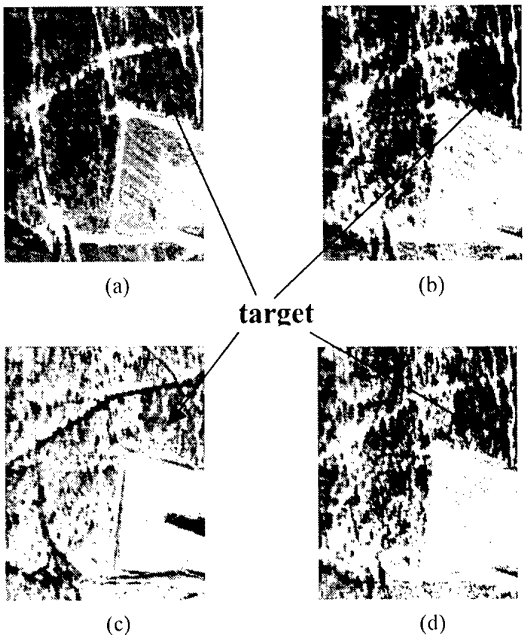


Figure 1. Single-band (a) short-wave infrared, (b) mid-wave infrared, (c) long-wave infrared, and (d) fused color was created by taking principle component direction of correlated thermal and visible pixel values as the luminance direction in a transformed space[23]. The airplane target is located in upper right quadrant.

Each background scene was 320 by 400 pixels (subtended 8.54^0 by 7.24^0 visual angle) with 50 percent of the stimuli containing a randomly placed airplane (subtended 0.1^0 by 0.1^0 visual angle). For each scene, the airplane target spectral characteristics was based on a measured target within the multispectral database. The long-wave target pixel spectral values were 255, mid-wave target pixel values were 73, and short-wave target pixel values were 114. The fused color scene spectral values were red=255, green=73, and blue=114. The achromatic fused images were spatially identical to the chromatic fused images; however the achromatic condition was employed to control for luminance effects. For each background scene, the target was present in 50 trials. The target location was generated by a random number generator and then inserted at that particular location. The target placement for each of the 50 locations was identical across the different background types.

The composite stimuli approach is to assign each pixel a color vector defined by the detected power in the registered three-band imagery[2]. Scatterplots (figure 2) of the image ensemble of colors frequently reveal pronounced anti-correlation

between short and long wavelengths, consistent with Kirchoffs law (reflective objects which appear bright in the short-wave infrared typically have low emissivity and appear dark in the long-wave infrared). For a given registered short- and long-wave infrared image pair, the principal component corresponds to the major axis–luminance channel. The orthogonal axis corresponds to the minor axis–color channel. The assignment of luminous intensities to the correlated component is straightforward, but the assignment of color to the uncorrelated features is not immediately obvious. The assignment of a pixel color is based on color opponency. By a-priori assigning one color to the image intensified (i^2) and it's color opponent to the infrared (*ir*), the resulting display shows two and only two opponent colors of various saturation. This makes an immediately intuitive representation as to which spectral bands dominant and by how much. It must be strongly emphasized that this system is mathematically incomplete to allow the perception of actual visible colors in the estimated reflectivity sense. Distinction between various vegetation, soil types, structures, water, and sky is based on coincident phenomenology in each spectral region, not by estimating a physical property such as emissivity.

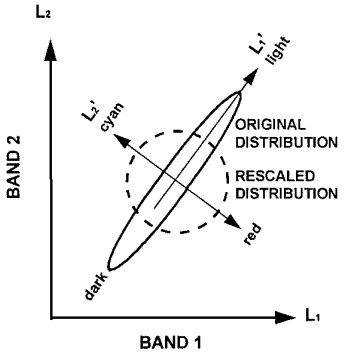


Figure 2. Color fusion algorithm technique. Two highly correlated bands will have a cigar-shaped distribution. The principal component direction (L_1') is the luminance channel and the orthogonal axis (L_2') is the chromatic channel. Increasing color contrast (dotted line) while retaining the luminance characteristics is achieved by re-scaling L_1'. In an actual sensor system, the principal component direction is based on the statistics of the scene (determined adaptively).

Procedure: Each subject participated in only one display format. Subjects were instructed to manually respond on a keyboard whether a target was present or absent within the scene. The four response categories were "1" = definitely no target to "4" = definitely a target. At the beginning of each trial, the subject fixated on a cross hair located in the center of the screen. The fixation cross was presented for 200 millisecond, immediately followed by a 1000 millisecond presentation of the experimental stimulus. The stimulus extinguished after the initial presentation or after the subject made a response, whichever came first. The next trial began approximately 1 second after the subject's preceding response. Accuracy was measured for each trial and no feedback was given for incorrect responses.

3.2. Matched Filter Analysis

In general the matched filter analysis paralleled behavioral testing described above. That is the same images and targets were used to generate numerical results. One additional step in the matched filter processing was to blur the image and target very slightly to take into account the modulation transfer function of the display and the human visual system. This was done using a narrow gaussian point spread function with a radius of one pixel. The actual computations were done using MATLAB™ software, which manipulates the image data in a matrix format. Single-band filters were derived using the smoothed 2-D power spectrum of the background and the target template with the target intensity identical to that used in the behavioral testing. The 3-D spatio-spectral (color) filter was derived by considering the target and the background as a 3-D space, with the third dimension being the spectral values of each pixel. In either case a multidimensional matched filter can be derived in the frequency domain using the expression,

$$H(\vec{k}) = c_0 \frac{S^*(\vec{k})}{W(\vec{k})} \exp(-i\vec{k}\vec{r}_0),$$

where $\boldsymbol{S}$ is the multidimensional signal representation, $\boldsymbol{W}$ is the multidimensional power spectral density. The spatial frequencies k_x and k_y are image coordinates indices in the frequency domain, and k_f is the spectral band index. The real space filter can be found by computing the inverse 3-D Fourier transform of $\boldsymbol{H}$,

$$h(x, y, \lambda) = F_{3-D}^{-1}\{H(k_x, k_y, k_f)\}$$

Processing the image with each respective filter is then done by convolving the filter with several hundred locations in the image. This is accomplished by multiplying filter values times corresponding pixel values aligned at each location. The summed values are stored for each position, giving an indication of false alarms (clutter leakage noise). These values are then compared to a second set of calculated totals produced by the same procedure, but with the target inserted at corresponding locations giving an indication of target detection. By comparing the signal-plus-noise values to the noise values for a given threshold value, false alarm and target detection probabilities can be calculated and displayed in the form of an empirical ROC plot.

3.3. Results

Signal detection theory distinguishes operator performance into two categories - sensitivity and response criterion or β[17]. Sensitivity is defined as the difference between the means of the signal plus noise and noise distributions. An observer's response criterion is independent of sensitivity. To calculate an observer's response criterion, β is equal to the ordinate of the signal plus noise distribution at criterion divided by the ordinate of noise distribution at criterion.

Both sensitivity and response criterion is derived from the probability of hits and probability of false alarms for each experimental condition. A ROC plot is a useful illustration of the relationship between sensitivity and response bias. The ROC curve plots on a single graph the joint value of

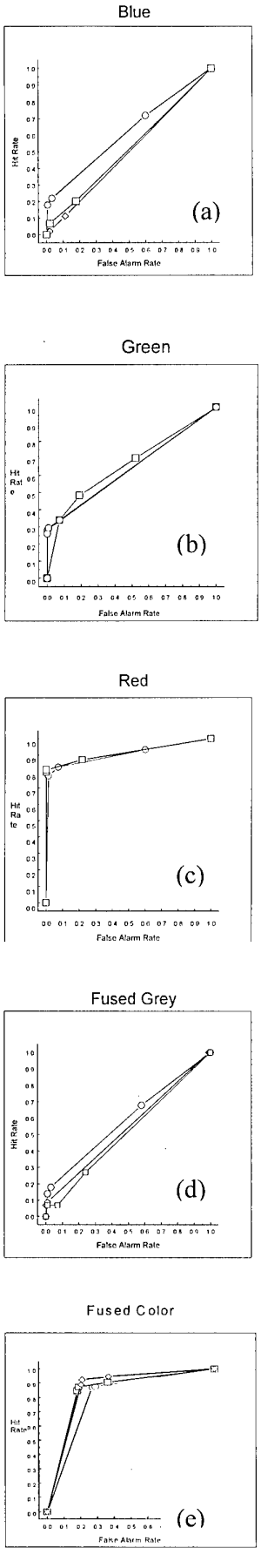


Figure 3 (left column). Human performance receiver operating characteristic (ROC) plots for fourteen subjects. Each format condition (a) short-wave infrared, (b) mid-wave infrared, (c) long-wave infrared, (d) monochrome fusion, and (e) color fusion had three subjects except for the gray fused condition had two subjects. Subjects within the fused color and long-wave infrared conditions had the highest sensitivity for detecting the target, while short-wave infrared and fused color near chance (d'=0).

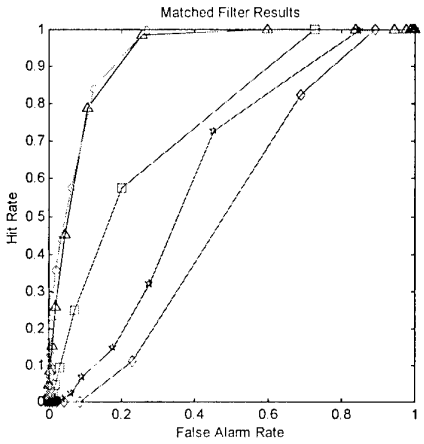


Figure 4. Matched filter ROC plot for the five different format types. Short-wave infrared = diamonds, mid-wave infrared = squares, long-wave infrared = triangles, fused-color = circles, fused monochrome = stars. Although the ROC sensitivities between each format are not quantitatively identical, the matched filter technique gives excellent qualitative agreement with the human performance tests.

probability of hits and probability of false alarms for each tested condition[24].

For this analysis, behavioral and matched filter ROC plots were compared across the five formats. Figure 3 illustrates the behavioral ROC plots across format types (blue = short-wave infrared; green = mid-wave infrared; red = long-wave infrared; fused gray = monochrome fused; and fused color = color fusion). The fused color and long-wave infrared formats had the highest sensitivity, while the short-wave infrared and fused gray sensor formats were near chance. The mid-wave sensor sensitivity was between the short- and long-wave sensor formats.

Figure 4 illustrates the matched filter ROC plot across format types. Again, the sensitivities show similar trends across format types. Moreover, the sensitivities between the matched filter and behavioral ROC plots are very similar. Therefore, the matched filter may be a viable alternative to human performance testing to assess operator detection performance.

4. EXPERIMENT 2

4.1. Introduction

Cognitive scientists record eye movements to understand cognitive processes that occur when an observer is searching for a target[25]. Eye movement data illustrates where and when the eye fixates within the scene; however, the data does not indicate what was processed. Rayner (1978) found that our eyes move three to four times per second while searching a scene. These saccadic eye movements enable the observer to extract important high spatial detail from each foveal fixation. Although there has been numerous eye movement studies investigating visual cognition, it is unclear what mechanisms control where and what the eye will fixate on next.

Biederman, Mezzanotte, and Rabinowitz (1982) found subjects extract information outside the fovea during scene perception. The parafovea and peripheral vision may extract certain features within a fixation; however, this information may or may not be identified. In order to integrate these parafovea cues into an identifiable object, a fixation is required[28]. To facilitate object identification, the more informative the scene the more likely the observer will fixate on those recognizable regions[28].

On the initial fixation, the observer will obtain a global snapshot of the scene. Next, low-level visual cues such as color, brightness, and contours will guide the observer's eye movements. Therefore, the level of informativeness within the picture will influence subjects' scene comprehension and object identification. It is hypothesized that a color-fused scene contains more informative features about the signal-to-noise ratio then an achromatic scene. Subjects initial eye movements will be guided to a color-fused target due to the target attributes. The achromatic target attributes will not contain enough informative information to capture the observers' visual attention. Furthermore, the eye movement results will correlate with the ROC plots. Targets embedded within the short- and mid-wave infrared scenes will require more saccades to identify the object, while the long-wave and fused conditions will be identified within the first couple fixations.

4.2. Methods

Subjects: Ten male military officers participated in this eye movement study. All subjects had normal (20/20), or corrected to normal, acuity and color vision. Subjects were naive to the purpose of the experiment and none had participated in previous visual search experiments. All subjects signed an informed consent and were briefed on the ethical conduct for subject participation in the Protection of Human Subjects[20].

Apparatus: Eye movements were recorded using an ISCAN, Inc. remote eye imaging system[29]. The eye tracker is a video-based system that uses an infrared camera to illuminate the eye and another camera to record the pupil to corneal reflection. The eye tracker then calculates the difference between the pupil and corneal reflection to indicate where the observer is fixating on the stimulus screen. The system operates at a sample rate of 60 Hz and the subject's visual point-of-regard may be determined with an accuracy of better than one degree over a +/- 25 degree horizontal to a +/- 20 degree vertical range.

Stimuli: Same as experiment 1.

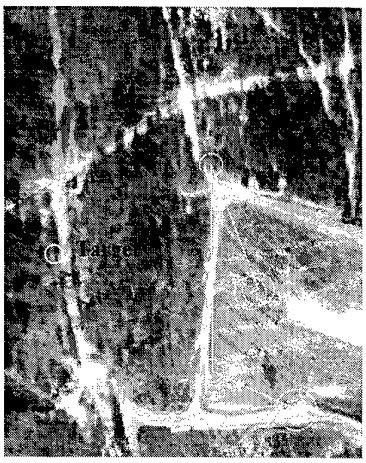


Figure 5a. Subject CH was not able to identify the short-wave infrared target. The subject's initial fixation provides enough global information about where to search within the scene, but the low target contrast does not have enough information to attract the visual system. This result correlates with the low sensitivity for the visual search task. Subjects' sensitivity was near chance (d'=0).

Figure 5b. Subject JL found the target after the first fixation. The color-fused target contained enough visual information to automatically guide the subject to the target location. Thus, the color-fused target had a high level of informativeness, which enabled the subject to identify the target with little effort. Again, this result correlates with the high sensitivity measure within the visual search experiment.

Procedure: Each subject participated in only one display format. At the beginning of the experimental session, the subject's head was placed in a chinrest positioned 1 meter from the stimulus monitor. The subject's right eye was then calibrated using a five-point calibration grid displayed on the

stimulus monitor. To maintain an accurate calibration between the eye tracker and the stimulus monitor, periodic five-point calibration checks were conducted throughout the experimental session.

At the beginning of each trial, the subject fixated in the center of the screen. Once the subject's eye was in the desired location, the experimenter initiated the trial. A fixation cross was presented for 200 milliseconds, immediately followed by a 5-second presentation of the experimental stimulus. Subjects were instructed to search for the target. Once the target was identified, the subject was to maintain fixation on the target until the stimulus extinguishes. There were 32 target trials and 16 noise trials presented for each format. Subjects' point-of-regard was recorded at 60Hz. No feedback on target identification accuracy was given.

4.3. Results

In each trial, the eye movement recording apparatus recorded the observer's fixation point at the rate of 60Hz, and hence a total of 300 data points were obtained per 5-second trial. An analysis software tool was subsequently used to analyze the data with the criterion for minimum fixation time at 40msec and the maximum horizontal and vertical deviation of the eyes at ±5 and ±3 pixels respectively. Thus, the number of fixations, the duration of each fixation, and the distance between fixations could be determined. These data were then tabulated to calculate the mean and the standard error mean values of the fixation duration, number of fixations, and scan path length.

Subjects within the short- and mid-wave infrared and gray fused conditions showed more fixations and longer scan-path lengths compared to the long-wave infrared and color-fused conditions. The long-wave and color-fused targets contained enough informative attributes to guide the subjects' eye movements to the desired location. Figure 5 illustrates a subject's search for a short-wave infrared and an another subject's search for a color fused target. The subject immediately identified the color-fused target, while the other subject was not able to find the short-wave infrared target. The subject within the short-wave infrared scene obtained enough global information within the initial fixation to search higher probability areas as to where the target may be located. However, the target's poor contrast inhibited the subject from identifying the location. Alternatively, the fused-color condition provided enough informative information within the first fixation to guide the subject to the target's location. The target's good spatial characteristics and large color contrast easily guided the subject to the appropriate critical region.

These results parallel the ROC results. The short- and mid-wave infrared and gray fused conditions low sensitivities match the eye scan data. Subjects within these conditions were not able to find the target within the first couple fixations which would indicate that their sensitivity should be low. Subjects within the color-fused and long-wave infrared conditions easily identified the target, which would indicate that their sensitivity should be high.

5. CONCLUSION

The purpose of this experiment was to compare matched filter analysis with human behavioral signal detection. The matched filter results illustrate that the different sensor format sensitivities are similar to the behavioral sensitivities. Although the ROC sensitivities between each format are not quantitatively identical, the matched filter technique gives excellent qualitative agreement with the human performance tests. Additional refinement of the matched filter should result in even better agreement. Ogawa (1997) found that the matched filter ROC was consistently superior to the behavioral ROC. His matched filter did not account for the human visual system inequalities. The gaussian blur was added to the filter to more accurately represent the human visual system resolution limit. The addition of gaussian blur to the filter caused our results to behave more similar to behavioral ROC as compared to Ogawa's results[30]. Additional refinement of the exact amount of gaussian blur to the matched filter should improve the correlation between the two ROC plots. The eye movement results illustrate that the eye was not able to identify the short- and mid-wave infrared and gray fused conditions as well as the color and long-wave infrared conditions. The color and long-wave infrared targets possessed important visual attributes that enabled the subject to identify the target with little to no effort. A surprising finding was the poor performance of the gray fused condition for both the signal detection and eye scan experiments. Subjects guided search for the target was not solely dependent upon spatial content; rather, visual search was mediated by both spatial and color target attributes. This finding indicates that color fusion is more appropriate for targeting applications than monochrome fusion. The color-fused target "pops-out" at the subject, which allows increased signal-to-noise sensitivity.

In summary, the matched filter technique may be a useful technique to predict human visual sensitivity for different sensor types by target characteristics. The matched filter technique will assist system engineers with a rough approximation of a human sensitivity to a target. This information could then be used for rapid prototyping of a system, enhance the predictability of existing electro-optical models, and provide a metric to test multi-spectral sensors. Additional tests will need to be conducted to test the robustness of the matched filter across different signal-to-noise ratios, terrain and target types, and various other atmospheric and illumination conditions. Finally, this matched filter will assist human factors testing by reducing the number of parameters needed to achieve the desired goal. Human factors testing will always be required, but at least the matched filter technique may provide the human factors group a better understanding of how the human will respond in the field.

6. ACKNOWLEDGEMENTS

Sponsored by DARPA's Integrated Imaging Sensors Program, Mr. Ray Balcerak is the program manager. A special thanks to James Lowell for his assistance in the eye movement data collection and analysis.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

7. REFERENCES

1. Krebs, W.K., Scribner, D.A., Miller, G.M., Ogawa, J.S., Schuler, J., "Beyond third generation: a sensor fusion targeting FLIR pod for the F/A-18", *Proceedings of the SPIE-Sensor Fusion: Architectures, Algorithms, and Applications II*, 3376, pp. 129–140, 1998.
2. Scribner, D.A., Satyshur, M.P., and Kruer, M.R., "Composite infrared color images and related processing", *Proceedings of the IRIS Specialty Group on Targets, Backgrounds, and Discrimination*, 1993.
3. U.S. Army Night Vision and Electronics Sensors Directorate, *AQUIRE range performance model for target acquisition systems*, Fort Belvoir, VA, 1995.
4. U.S. Army Night Vision and Electronics Sensors Directorate, *FLIR92 thermal imaging systems performance model.* Fort Belvoir, VA, 1993.
5. McDaniel, R., Scribner, D., Krebs, W., Warren, P., Ockman, N., McCarley, J., "Image fusion for tactical applications", *Proceedings of the SPIE - Infrared Technology and Applications XXIV*, 3436, pp. 685–695, 1998.
6. Toet, A., van Ruyven, L. J., & Valeton, J. M., "Merging thermal and visual images by a contrast pyramid", *Optical Engineering*, 28, pp. 789–792. 1989.
7. Palmer, J., Ryan, D., Tinkler, R., Creswick, H. , "Assessment of image fusion in a night pilotage system", *NATO AC/243 Panel 3/4 Symposium on Multisensors and Sensor Fusion*, Brussels, Belgium, 1993.
8. Toet, A., & Walraven, J., "New false color mapping for image fusion", *Optical Engineering*, 35, pp. 650–658, 1996.
9. Therrien, C.W., Scrofani, J., and Krebs, W.K., "An adaptive technique for the enhanced fusion of low-light visible with uncooled thermal infrared imagery", Proceedings *of the IEEE: International Conference on Imaging Processing*, pp. 405-408, 1997.
10. Waxman, A. M., Gove, A. N., Fay, D. A., Racamato, J. P., Carrick, J. E., Seibert, M. C., & Savoye, E. D., "Color night vision: Opponent processing in the fusion of visible and IR imagery", *Neural Networks*, 10, pp. 1–6, 1997.
11. Essock, E.A., Sinai, M.J., McCarley, J.S., Krebs, W.K., and DeFord, J. K., "Perceptual Ability with Real-World Nighttime Scenes: Image-Intensified, Infrared and Fused-Color Imagery", *Human Factors*, (in press).
12. Sampson, M.T., Krebs, W.K., Scribner, D.A., and Essock, E.A., "Visual search in natural (visible, infrared, and fused visible and infrared) stimuli", *Investigative Ophthalmology and Visual Science*, (SUPPL) 36, 1362, FT Lauderdale, FL, 1996.
13. Steele, P.M. and Perconti, P., "Part task investigation of multispectral fusion using gray scale and synthetic color night vision sensor imagery for helicopter pilotage", *Proceedings of the SPIE Conference on Aerospace/ Defense Sensing, Simulation, and Controls,* 3062, pp. 88–100, 1997.
14. McCarley, J.S., Krebs, W.K., Essock, E.A., and Sinai, J.S., "Multidimensional scaling of single-band and sensor-fused dual-band imagery", (in review).
15. Sinai, M.S, McCarley, J.S., and Krebs, W.K., "A comparison of sensor fusion and single band sensors in the recognition of nighttime scenes", *IRIS Passive Sensors*, pp. 1–9, Monterey, CA, 1999.
16. Sinai, M.J, McCarley, J.S., Krebs, W.K, and Essock, E.A., "Psychophysical comparisons of single- and dual-band fused imagery", *Proceedings of the SPIE-Synthetic Advanced Vision,* 3691, pp. 1–8, 1999.
17. Tanner, W.P., & Swets, J.A., "A decision-making theory of visual detection", *Psychological Review*, pp. 401–409, 1954.
18. Green, D.M., & Swets, J.A., *Signal detection theory and psychophysics*, Wiley, New York, USA, 1966.
19. Pratt, W. K., *Digital Image Processing*, John Wiley & Sons, Inc, New York, 1991.
20. Eckstein M.P., & Whiting, J.S., " Visual signal detection in structured backgrounds. I. Effect of number ofpossible spatial locations and signal contrast", *Journal of the Optical Society of America (A)*, 13(9), pp. 1777-87, 1996.
21. Department of the Navy, *Protection of human subjects*, (SECNAV Instruction 3900.39B). Washington, D.C.: Chief of Naval Operations OP-098, 1984.
22. Swift, D.J., Panish, S. and Hippensteel, B., " The use of VisionWorks in visual psychophysics research," *Spatial Vision*, 10, pp. 471–477, 1997.
23. Scribner, D.A., Warren, P., Schuler, J., Satyshur, M., and Kruer, M., "Infrared color vision: an approach to sensor fusion", *Optics and Photonics News*, 8, 27–32, 1998.
24. Gescheider, G.A., *Psychophysics method, theory, and publication, 2nd edition.* Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey, 1985.
25. Rayner, K., "Eye movements and visual cognition: Introduction", In K. Rayner (Ed.), *Eye Movements and Visual Cognition Scene Perception and Reading*, Springer-Verlag, New York, 1992.
26. Rayner, K., "Eye movements in reading and information processing", *Psychological Bulletin*, 85, pp. 618–660, 1978.
27. Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C., "Scene perception: Detecting and judging objects undergoing violation", *Cognitive Psychology*, 14, pp. 143–177, 1982.
28. Rayner, K. & Pollatsek, A., "Eye movements and scene perception", *Canadian Journal of Psychology*, 46(3), pp. 342–376, 1992.
29. Razdan, R. & Kielar, A., "Eye tracking for man/machine interfaces", *Sensors*, September 1988.
30. Ogawa, J.S., *Evaluating color fused image performance estimators*, Master of Science in Operations Research, Naval Postgraduate School, 1997.

DETECTION OF LOW-CONTRAST MOVING TARGETS

John P. Mazz
Regina W. Kistner
William T. Pibil
U.S. Army Materiel Systems Analysis Activity
392 Hopkins Road
Aberdeen Proving Ground, Maryland 21005-5071
The United States of America
E-mail: mazz@arl.mil

1. SUMMARY

The U.S. Army Materiel Systems Analysis Activity (USAMSAA) designed a perception experiment to assess the influence of target angular velocity on the detectability of low to moderate contrast targets. The Moving Target Experiment II (MTE II) was designed to be representative of search with the unaided eye. Target angular velocity, range, contrast, and background were varied. Targets with near-equal contrast at identical range and angular velocity yielded widely different probabilities of detection. However, within a specific background region, contrast had a significant impact. This localized impact of target contrast indicates that further improvements in search and target acquisition modeling requires the evaluation of scene-content's impact on target detection (i.e., what about the scene leads an observer to the vicinity of the target.) For low-contrast targets, scene content has even greater impact on detection.

The U.S. Army's standard methodology for representing search and target acquisition in combat models is the ACQUIRE model. Current implementations of ACQUIRE utilize the "two-thirds rule" to represent the detection of all moving targets regardless of angular velocity. The n50 for the detection of moving targets is simply 2/3 of the n50 used to represent the detection of stationary targets. Results of the MTE II and other experiments indicate that the appropriate ratio of moving-to-stationary n50 decreases as a function of angular velocity. A ratio of 2/3 equates to an angular velocity of 1 milli-radian/sec and a ratio of 1/3 equates to an angular velocity of 3.3 milli-radians /sec.

Keywords: Detection, moving targets, search, target acquisition, false targets, n50, Johnson criteria, perception experiment, low contrast, ACQUIRE model

2. INTRODUCTION

The U.S. Army's standard methodology for modeling man-in-the-loop target acquisition performance is the ACQUIRE Model developed by the U.S. Army Communications and Electronics Command Research Development and Engineering Center's Night Vision and Electronic Sensors Directorate (NVESD).[1] ACQUIRE predicts the probability of detection (Pd) as a function of Minimum Resolvable Contrast (MRC); target size, range and contrast; and an observer task parameter called n50. The MRC provides the minimum contrast at which a specific spatial frequency (cycles per milli-radian) can be resolved by an observer. The n50 parameter (also known as the Johnson criterion) is defined as the number of cycles across the target (at the maximum resolvable spatial frequency) required for 50 percent of the observer population to detect the target. The Johnson criterion is also used to represent higher levels of target discrimination such as recognition and identification; however, this effort is concerned only with detection.

In current implementations of ACQUIRE, the value of n50 for moving targets is set to two-thirds the value of n50 for detection of stationary targets. This two-thirds value was derived from results of the Moving Target Experiment I (MTE I) [2]. Prior to MTE I, a n50 value of 0.5 was used to represent detection of moving targets.

Modeling the detection of moving targets with a single n50 value is rather simplistic. Under this approach, the angular velocity of the target is not taken into consideration. Slow moving targets are equally as detectable as fast moving targets. This approach may be adequate for most direct-fire battle scenarios; but for the assessment of the value of low-contrast in the scout mission, this approach is woefully inadequate.

Results of MTE I gave indications that, as expected, the n50 for moving targets decreases with increasing angular velocity. Unfortunately, since ground speed was the primary parameter representing velocity at each range, sample sizes with respect to angular velocity were small. MTE I was also limited to foveal detection, no search was involved. Since search is a significant aspect to most tactically realistic scenarios, MTE II was designed as a search experiment to collect statistically significant samples with respect to angular velocity. As a result of the MTE II experiment, we hope to further refine the n50 values used for moving targets.

The MTE II laboratory perception experiment was designed and analyzed by the U.S. Army Materiel Systems Analysis Activity (AMSAA) under the auspices of the Joint Technical Coordinating Group for Munitions Effectiveness (JTCG/ME). The test was conducted by the U.S. Army Tank-Automotive and Armaments Command Research, Development and Engineering Center (TARDEC) at their Perception Laboratory in Warren, Michigan in 1998.

3. EXPERIMENT DESIGN

The experimental design consists of four primary parameters of interest: background, target size (simulated range), target contrast, and velocity. In all, 565 sequences were presented to each of the observers -- 80 percent with targets and 20 percent without targets. The 80 percent with targets consisted of a full-factorial experimental design of the four parameters. The primary purpose of the no-target trials was to discourage guessing by allowing the possibility of no target present. Furthermore, these no-target trials provide a means of quantifying an observer's willingness to guess.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

3.1. Stimuli

The experimental stimuli were created from 35mm visual imagery taken during a field test exercise. This field exercise was conducted in a desert environment. Five images were chosen, each having different clutter and target contrast levels. These images are referred to as background images 4, 7, 12, 14, and 15. The target in each image was a side aspect military vehicle. All five images were manually adjusted to represent three conditions: stationary target, moving target, and no-target. The targets were separated from the background images and the backgrounds were "fixed" by substituting appropriate background in the area vacated by the target.

The .AVI movies were developed using commercial software to place the targets into the background scenes. The target starting and ending locations were calculated in order to represent the appropriate velocities for a specified range. The targets moved perpendicular to the observer's line-of-sight in either the left or right direction. For the no-target image, no target was inserted into the scene.

Adobe Premiere software was used to create the motion sequences. Motion sequences were generated with lateral ground speeds dependent upon the simulated range. For the 750 meter range the four speeds were: 0, 2.5, 6.25 and 10 kilometers per hour (kph); while the 1500 meter range speeds were 0, 5, 12.5 and 20 kph; and the 3000 meter speeds were 0, 10, 25, and 40 kph. The four speeds at each range are represented by a single set of angular velocities of 0, 0.9, 2.3 and 3.7 milli-radians per second.

The targets started at either a left, central or right grid for a given range. The direction for the centrally located targets was determined randomly, while targets in locations on the right moved left and targets in locations on the left moved right. The target images were scaled to represent different ranges and then placed into the location in the background that corresponded to that range. The observer was seated 1.3 meters from a nominal 17 inch monitor and the target size was configured to represent three simulated ranges: 750, 1500 and 3000 meters. The software was also used to create stationary and motion sequences for targets with reduced contrast. The original image represented a high contrast target. The brightness level (for each of the Red, Green and Blue color spectra) of the target was decreased to represent lower contrasts. This process was conducted twice, yielding three contrast levels for each image.

The stimuli were constructed to represent an observer viewing the scene with the unaided eye. Each image sequence represented pure motion. There were no secondary environmental effects such as dust or motion-of-vegetation represented.

3.2. Conduct

The perception experiment was conducted by the TARDEC at their Perception Laboratory in Warren, Michigan. The computer display used was a Panasonic PanaSync/Pro P17 monitor. Observers were required to sit 1.3 meters from the screen to accurately simulate the ranges of interest. Each observer went through a training session. The training included written instructions read together with the test controller and an opportunity for the subject to go through the software and images, until the subject was thoroughly familiar with the images, targets, and test procedures.

A total of 22 observers participated in the perception experiment. The observers were recruited by a research firm and were reimbursed for their participation. Each subject had some military experience, either active Army, Reserves or the National Guard. The subjects were between 25 and 45 years of age with normal/corrected 20/20 vision. Prior to participating in the experiment, each subject was screened for vision abnormalities using a Snellen chart and Ishihara color plate book. Each observer was presented with the entire set of 565 images. The order of presentation was randomized for each observer. Each image sequence lasted up to 9 seconds. Upon detecting the target, the observer clicked the computer mouse button and the target motion stopped. The observer was then required to use the mouse to click on the location of the target on the screen. The subjects were instructed that they could rest at any time during the test, by hitting the 'O' key or by telling the test controller to pause the test. The observer could then resume the test when ready.

The experiment lasted approximately 1 to 2 hours for each subject. No feedback was provided to the observers during the experiment; however, upon completion of the experiment, overall performance was provided to the observer upon request.

4. ANALYSIS OF RESULTS

The purpose of this experiment was to investigate the effects of target motion on the probability of target detection (Pd). Target detection was scored in the following manner. During the experimental trials, each observer was asked to locate the target he detected with crosshairs controlled by the computer mouse. A scoring box was created that was centered on the target and twice the length and width of the target. If the crosshair location given by the observer was within this box, the observer was scored to have a correct detection; otherwise, he was scored to have a false detection. [This scoring process was not perfect. It may score a false target as true since the scoring box is bigger than the target and a true target as false since the motion stopped after the observer indicated detection but before he indicated the target location.] Pd is estimated as the number of observers who detect divided the total number of observers (22).

The main parameters varied in this experiment were target angular velocity (0.0, 0.9, 2.3, and 3.7 milli-radians per second), simulated target range (750, 1500, and 3000 meters), and target contrast (original and attempted 50% and 75% reductions). Figure 1 shows the average Pd, collapsed over contrast, as a function of target angular velocity and range. As expected, Pd is strongly effected by both angular velocity and range. Pd increases with both increasing angular velocity and decreasing range. The corresponding ground speeds in kilometers per hour (kph) for the simulated lateral motion (perpendicular to the observer's line-of-sight) are presented in Table 1. It is interesting to note that even with a ground speed as high as 40 kph, the Pd is still below 0.5 at 3km. Although increased angular velocities would likely increase this Pd, the practicality of exceeding 40 kph as a cross-country ground speed is unlikely. It should also be noted that as you transition from lateral to radial motion, angular velocity decreases.

Table 1 Simulated Ground Speeds in kph

Range \ Velocity	0.9 mr/sec	2.3 mr/sec	3.7 mr/sec
750 m	2.5	6.25	10
1500 m	5	12.5	20
3000 m	10	25	40

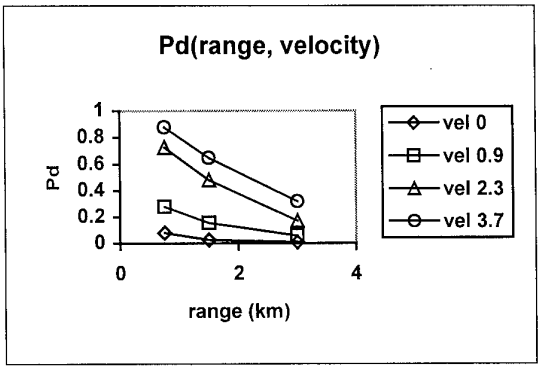


Figure 1 Probability of detection

In the analysis of target detection, it is important to investigate false detections. It provides insight to an observer's criterion (i.e., the trade-off between increased Pd and more false targets). Knowledge of false target performance is essential when comparing the results of different experiments as well as the results from different observers. Figure 2 shows the false target results of this experiment. The probability of selecting a false target on the 20% of trials containing no-targets was approximately 0.24. The four curves represent the probability of selecting a false target over a true target when the true target had specific values of range and angular velocity. Since these were single target trials, the observer could not detect both a true and a false target on the same trial; therefore, the sum of Pd and the probability of false target for a particular trial is always less than or equal to one. As range to the true target decreases or the angular velocity of the true target increases, the probability of false target decreases. Since the only motion in these scenes involved the target, it is safe to assume that all false targets were perceived as stationary by the observer. One anomaly in Figure 2 is that the average probability of false target (collapsed over target contrast) for the 0.9 mr/sec angular velocity is greater than that for the stationary trials. This difference is not statistically significant at the alpha=0.05 level (p-value=0.268, one-sided Sign Test).

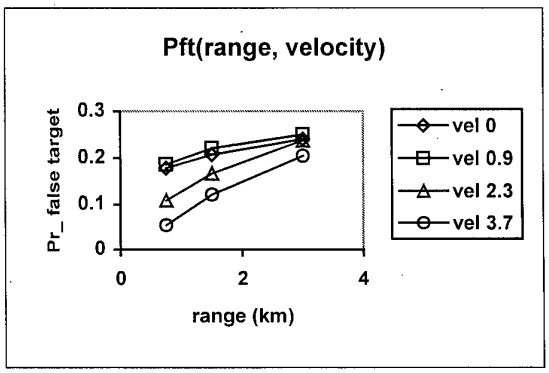


Figure 2 Probability of false target

The third main parameter varied in this experiment was target contrast – the original dark target and an attempt to reduce contrast by 50% and 75%. The chief measure of contrast used in this analysis was luminance (not color) contrast based on the area-weighted average of target and background image-pixel luminance-grayscale values. [In general, the area-weighted average contrast is a non-unique metric for real-world scenes. Its value varies depending on what part or how much of the background is included in its calculation.] For simplicity, an average contrast value is used to represent the target contrast for the entire 9-second motion sequence. For the initial look at the effects of target contrast, the Pd was averaged within the following three contrast bins: 0.0 to 0.2, 0.2 to 0.4, and greater than 0.4. Pd's were placed into bins based on the absolute value of contrast. The majority of the contrasts were negative. Figure 3 shows Pd as a function of contrast bin, range, and angular velocity. Figure 3 is separated into three range regions; the x-axis going from an absolute contrast of 0.0 to 1.0 within each of these regions. Contrast seems to have a much milder effect on Pd than range or velocity. This is confirmed by the regression analysis presented in Table 2 where the inclusion of contrast explains only an additional 1% of the variation in Pd exhibited during this experiment.

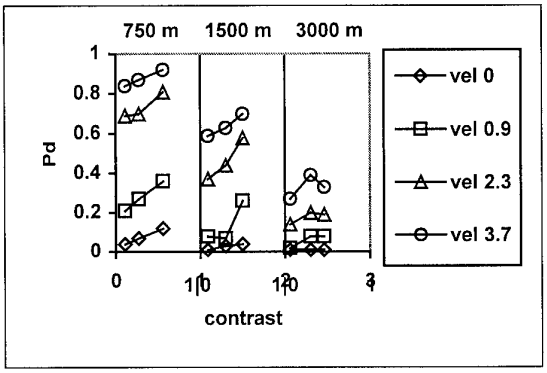


Figure 3 Pd versus range, velocity, and contrast

Table 2 Regression Analysis

Regression Variables	Percent Variation in Pd Accounted For
Angular Velocity only	48%
Range only	16%
Contrast only	3.6%
Velocity & Range	64%
Velocity & Range & Contrast	65%

One reason for the muted effect of target contrast is that the experiment involved search of a complex scene. Where an observer looks is determined by scene content and not by target contrast. However, once looking in the vicinity of the target, contrast has a greater impact as illustrated in Figure 4. Figure 4 shows the Pd results for the individual trials utilizing background image #14. The symbols in Figure 4 can be identified by the legend in Figure 3. Each vertical region represents one of nine potential target locations (listed at top of Figure 4). Locations 1, 2, and 3 are the top row (3000m); locations 4, 5, and 6 are the center row (1500m); and locations 7, 8, and 9 are the bottom row (750m) of the image. Locations 1, 4, and 7 are on the left; locations 2, 5, and 8 are in the center; and locations 3, 6, and 9 are on the right side of the image. In general, Figure 4 shows that as contrast increases so does Pd. However, comparing the 3 locations associated with a specific range, it can be seen that the Pd results are widely different. For example, although the highest contrast in locations 1, 2, and 3 is approximately 0.33, Pd's are much higher in location 1 than in locations 2 and 3. Similarly, the Pd's associated with the 0.9 mr/sec velocity in location 9 are much lower than the

Pd's in locations 7 and 8 even though contrast is nearly identical. One possible explanation for the above effect of identical contrasts yielding widely divergent Pd's is that the observers were less likely to look in the locations with lower Pd's. This localized impact of target contrast indicates that further improvements in search and target acquisition modeling requires the evaluation of scene-content's impact on target detection (i.e., what is it about the scene that leads an observer to the vicinity of the target.) For low-contrast targets, scene content has an even greater impact on detection.

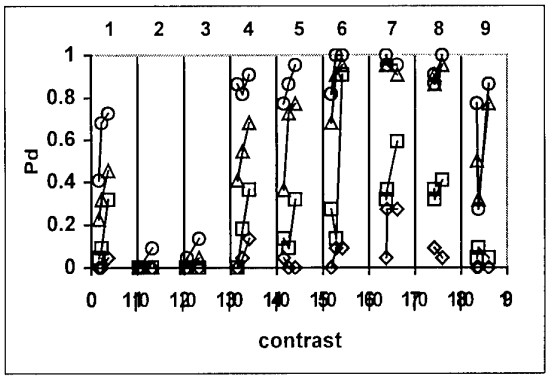


Figure 4 Pd results for background # 14

5. MODELING IMPLICATIONS

The U.S. Army's standard methodology for representing search and target acquisition in combat models is the ACQUIRE model[1]. Current implementations of ACQUIRE utilize the "two-thirds rule" to represent the detection of all moving targets regardless of angular velocity. The n50 for the detection of moving targets is simply 2/3 of the n50 used to represent the detection of stationary targets. Admittedly, this is an oversimplification; however there has been little concrete data on which to base a more complex solution.

The MTE II experiment indicates that the ratio between moving and stationary n50 decreases with increasing angular velocity. Table 3 presents the n50's which best represent the experimental results for each angular velocity. The 2/3 rule seems appropriate for the lowest angular velocity but not for the higher angular velocities. The n50 values marked by the asterisks have been adjusted to account for an incomplete spectrum of probabilities. (i.e., 97% of the observed Pd's were less than 0.2 for the 0 mr/sec and 0.6 for the 0.9 mr/sec angular velocities.)

Table 3 n50's

Angular Velocity	n50	n50 Ratio (moving/stationary)
0.0 mr/sec	5.7*	1
0.9 mr/sec	4.1*	0.72
2.3 mr/sec	2.5	0.42
3.7 mr/sec	1.9	0.33

Figure 5 plots the n50 results for the MTE II experiment along with the results of two previous experiments – MTE I and Summer 94. MTE I was a foveal vision experiment. When present, all targets appeared in the center of the image. No search was involved. The approach to simulating an increase in range was to shrink the image. At the 1500 m target range the image subtended 6° horizontally, and at 6000 m the image subtended 1.5°. The shrinking images in MTE I had a confounding effect on n50 (n50 decreased with shrinking image). Since the only angular velocities in common between MTE I and II occurred at 1500 m and because of the confounding effects of shrinking images, only the 1500 m MTE I n50's are presented in Figure 5. Further information on MTE I can be found in reference 2.[2]

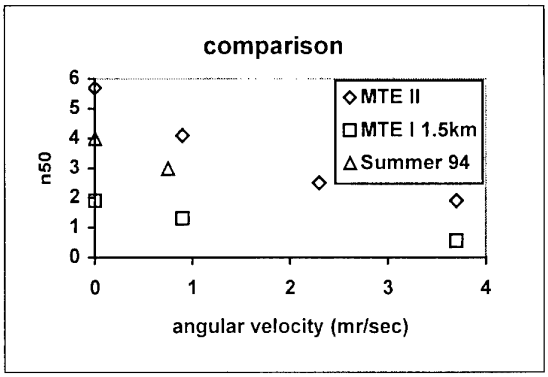


Figure 5 n50 versus velocity and test

The Summer 94 field experiment was the source of the background and target imagery used in both MTE I and MTE II. The moving and stationary target trials were conducted on separate days.

Table 4 provides a comparison of key parameters for the three experiments. The differences in these parameters may explain the wide variability in the resulting n50's. The observers were basically performing different tasks in each of these experiments resulting in different n50's. Reference 3 provides further information on how the n50 for stationary targets varies in relation to such factors as clutter, false targets, sensor resolution, and observer task.

Table 4. Comparison of Test Parameters

	MTE II	MTE I	Summer 94
Search	6° image	6° to 1.5° image	120° sector
Experiment Type	Lab	Lab	Field
Time Limit (seconds)	9	3	60 (moving) 180 (stationary)
Range (km)	0.75-3.0	1.5-6.0	0.5-1.5
Average False Targets per Observer per Trial	0.19	0.16	0.008 (mov) 0.08 (stat)
Average Pd	0.32	0.35	0.60 (mov) 0.47 (stat 60 sec) 0.64 (stat 180 sec)

Even though the resulting n50's were different, the ratios of moving target to stationary target n50 were remarkably similar as illustrated in Figure 6. The ratio decreases as angular velocity increases. The following equation produced the line that fits the data in Figure 6.

$$\text{Ratio} = \exp(-0.4 * \text{ang.vel.}^{0.85}) \quad (1)$$

Equation (1) indicates that the current "2/3 ratio rule" equates to an angular velocity of 1.0 mr/sec. However, the ratio reduces to 1/3 at an angular velocity of 3.3 mr/sec. Since a non-zero lower limit on the ratio likely exists, care should be taken when extrapolating equation (1) beyond an angular velocity of 3.7 mr/sec.

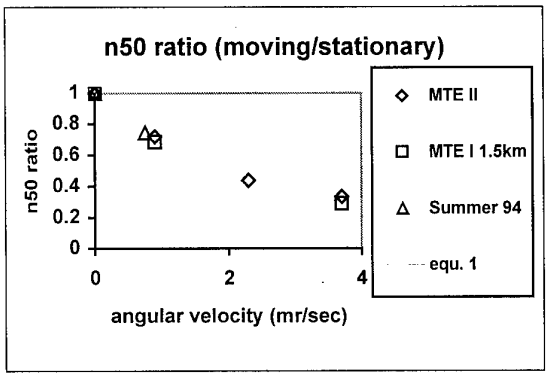


Figure 6 n50 ratios (moving/stationary)

It should be noted that equation (1) has been developed from a limited set of data and may have limited applicability. The experimental data used represents unaided eye performance at target ranges of 3000m or less. Additional experiments are required to investigate the applicability of equation (1) to magnified optics and thermal sensors.

6. FURTHER ANALYSIS

Analysis of the MTE II is ongoing. Additional areas of investigation include: variations in contrast as the target moves (does greater variation lead to higher Pd or faster detection?), time to detect (how does time to detect correlate with the peaks and nulls in contrast and background variations?), background characteristics (what is it about one region of the background that makes a 0.3 target contrast more detectable than in another region of the background of equal range?).

7. ACKNOWLEDGEMENTS

The authors would like to recognize the following individuals and organizations for their contributions to the conduct and analysis of the MTE II experiment:

Dr. Tom Meitzler, U.S. Army TARDEC and the TARDEC Perception Laboratory Team conducted the experiment.

Scott Schoeb and Lilly Harrington, USAMSAA constructed the experimental stimuli.

The Joint Technical Coordinating Group for Munitions Effectiveness (JTCG/ME) funded the design and conduct of the experiment.

8. REFERENCES

1. Howe, J.D., "Electro-Optical Imaging Systems Performance Prediction," *The Infrared and Electro-Optical Systems Handbook*, Volume 4, Chapter 2, Infrared Information Analysis Center, Michigan, and SPIE Optical Engineering Press, Washington, 1993.
2. Kistner, R.W., Pibil, W.T., Meitzler, T.A., "AMSAA/TARDEC Moving Target Perception Experiment and Analysis," *Proceedings of the Eighth Annual Ground Target Modeling and Validation Conference*, pp. 201-205, Signature Research Inc., Calumet, Michigan, USA, August 1997.
3. Mazz, J.P., "ACQUIRE Model: Variability in n50 Analysis," *Proceedings of the ninth Annual Ground Target Modeling and Validation Conference*, pp. ???, Signature Research Inc., Calumet, Michigan, USA, August 1998. (publication pending)

VALIDATION AND VERIFICATION OF A VISUAL MODEL FOR CENTRAL AND PERIPHERAL VISION

Eli Peli
Schepens Eye Research Institute, Harvard Medical School,
20 Staniford St., Boston MA 02114 U.S.A.
Phone (617) 912-2597; Fax (617) 912-0111
E-mail: eli@vision.eri.harvard.edu

and

George A. Geri
Visual Research Lab, Raytheon Training Inc.
6030 South Kent, Mesa, Arizona, 85212 U.S.A.

1. SUMMARY

Many computational visual models use the contrast sensitivity function (CSF) to represent certain visual characteristics of the observer. In addition, these models are often implemented using a multi-scale, band-limited representation of image contrast. The purpose of the present study was to evaluate a previously described visual model (Peli, *JOSA A*, **7**, 2030, 1990) by comparing the appearance of an image viewed at various distances with simulations of that image corresponding to the same distances generated with the model. Among the unique characteristics of this model are that it applies a threshold (i.e. nonlinear) CSF and a locally normalized, band-limited contrast. Since CSFs can vary substantially depending both on the stimuli and the testing method used to measure them, the model was evaluated using several CSFs. The model was also evaluated for both central images, extending to 2° eccentricity, and peripheral images, extending from 8° to 32° eccentricity. Changes in the images with eccentricity were modeled by a single parameter. For the central (2°) stimuli, the CSF obtained with 1-octave Gabor stimuli and a contrast detection task provided better simulations than the other CSFs tested. In addition, data obtained using both lower and higher contrast versions of the same images verified the CSF over a wide range of frequencies and indicated that the model was sensitive to small variations in the chosen CSF. For the peripheral (6.4°-32°) stimuli, the same 1-octave, detection CSF was found to provide the best simulation. In general, the model suggested by Peli (1990) performed well for both the central and peripheral visual targets, suggesting that the use of a nonlinear CSF and locally normalized contrast are valid. Further, the performance of the model for the peripheral stimuli suggests that, at least for the simple discrimination task used here, differences in image detail across wide-field images can be modeled using a single eccentricity-dependent parameter in addition to the foveal CSF.

Keywords: vision models, simulation, contrast, CSF, wide field, peripheral retina, nonlinear processing

2. INTRODUCTION

Simulations of the appearance of visual images and scenes have been studied in many areas of visual science [1-4], and engineering [5-7]. The simulations are usually generated using a computational vision model. One such multi-scale model of spatial vision was used to calculate local band-limited contrast in complex images [8]. This contrast measure, together with observers' contrast sensitivity functions (CSFs), expressed as thresholds, was used to simulate the appearance of images to observers with normal vision [8] and low vision [7]. Others have applied the same concept of local band-limited contrast with small variations [6, 9, 10]. The local band limited contrast model was also used to simulate the appearance of images presented to the peripheral retina [11] using the CSF measured at various retinal eccentricities. Validation of the simulations and the underlying vision models is crucial for such applications.

We summarize here the results of tests of both central (2°) and peripheral (6.4°-32°) visual models performed using simulations of complex images. The peripheral model is identical to the foveal model except for the addition of a single parameter representing the change in the contrast detection threshold across the retina. In the foveal study, observers were asked to discriminate an original image from a simulation of the original as viewed from various distances. The distance at which discrimination performance was at threshold was compared with the simulated observation distance, and was found to be the same. In the peripheral study, we modeled the change in contrast sensitivity with eccentricity, and compared the data to those obtained using simpler stimuli as reported in the literature.

While it appears that methodological differences may account for the variability of the CSF data in the literature, we do not know yet which method should be used to obtain the CSF that is most appropriate for simulating the appearance of complex images in the context of pyramidal, multi-scale vision models. Therefore, we have compared the appearance of complex images simulated from CSFs obtained using test gratings whose spatial extent was determined by either a constant-width (square) or a variable-width (1-octave gaussian) window. Further, we compared the appearance of complex images simulated from CSFs obtained using either pattern detection or an orientation detection task.

In applying the vision model to simulations or other applications one needs to consider both the object's contrast spectrum, computed in terms of cycles/object or

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

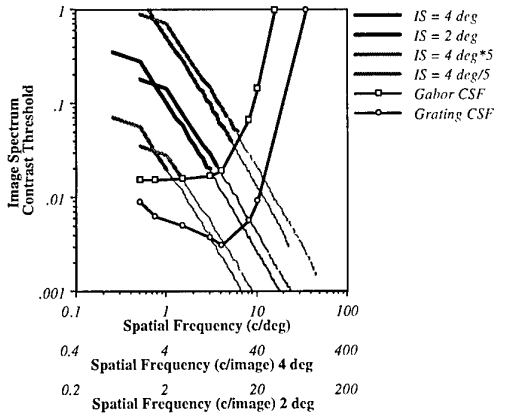


Fig. 1. The interaction, from two different observation distances, of image spatial frequency content at different image contrasts (amplitudes) with two example CSFs. The thick black line represents a typical (1/f) image spectrum (IS) for a 4° image. The part of the IS below the upper CSF (pattern detection threshold obtained with Gabor stimuli) will not be detectable (shown as thin black line). A change in observation distance that decreases the image to 2°, shifts the IS along the 1/f line (second thick black). At the new distance, lower object frequencies are removed by the observer's CSF but essentially the same retinal frequencies are involved. The gray pairs of curves represent the spectra of images with increased (upper pair) and decreased contrast, which allow testing of other parts of the CSF since they intersect it at higher and lower retinal frequencies, respectively.

cycles/image, and the observer's CSF expressed in terms of cycles/degree (c/deg.). To derive the object's spectrum at the retina (see Fig. 1, solid black curve), the distance of the object from the observer needs to be known. Any information in the image that falls below the observer's threshold (i.e., below the point at which the contrast threshold curve intersects the image spectrum curve) is not visible to the observer, and should not be included in the simulation. This is illustrated in Fig. 1 by the spectrum lines turning to thin lines at the values that are below threshold. If the original and simulated images are viewed from the simulated distance or farther, they should be indistinguishable, as the image information that is below threshold in the original would not be used in the simulation. However, if the original and simulation are viewed from a closer distance, the difference in content between the original and the simulation should be visible.

When the distance between an object and an observer increases, the retinal size of the object decreases and its retinal spatial frequencies increase. It was previously thought by us [12] and others [13] that this change results in a shift of the object's spatial spectrum to the right along the spatial frequency axis (see Fig.1). However, as Brady and Field [14] pointed out, the spectrum actually shifts both to the right (higher frequencies) and down (lower contrast) sliding along a line with slope = -1.0. Most natural images have a spatial frequency spectra that behaves as 1/f (i.e., have slope = -1.0). Thus, a change in object size causes the spectrum to "slide along itself" (Fig. 1, pair of thick black lines). As a result, the spectrum of the farther image intersects the CSF curve at essentially the same retinal frequency. Only the mapping of the relevant object frequencies to the retinal frequencies changes. Therefore, the experiments reported by Peli [15] have probed only a very limited range of spatial frequencies in the CSF. To further examine the CSF, one needs to use images whose spectra intersect the CSF at other frequencies. This was achieved here by using higher and lower contrast versions of the same images, as illustrated schematically in Fig. 1. As a practical matter, we actually increased or decreased the amplitude of the images, not their contrast. This operation in which the image dc value is subtracted and the remaining values are scaled up or down is frequently referred to as a contrast increase or decrease. As pointed by Peli [8], however, the changes in contrast are equivalent to changes in amplitude only when the local luminance is equal to the mean luminance. We use the term contrast change here for consistency with previous work. The higher- and lower-contrast image spectra intersect the threshold CSF curve at higher and lower spatial frequencies, respectively, and thus can be used to test the CSF at those additional frequencies.

The issues discussed above suggest that, despite the fact that complex images were used, one limitation of our previous work [12, 15, 16] is that the validity of the chosen CSF was tested at one spatial frequency only. In the work described here, we sought to expand this investigation to a wider range of retinal spatial frequencies by using as stimuli images scaled in contrast over a correspondingly wider range. As was described above, the lower contrast images effectively were used to test the lower retinal spatial frequencies, while the higher contrast images were used to test higher spatial frequencies.

3 FOVEAL VISION MODEL

3.1 Methods

We tested our vision model by comparing an original image with a model simulation of how the original would appear from various distances. If the model is valid, the original and simulated images should be indistinguishable at distances equal to or greater than the distance used for the simulation. Likewise, the two images should be progressively easier to distinguish at distances shorter than the simulated distance.

Observers viewed image pairs from various distances presented in a forced-choice procedure. Simulated test images were obtained using each observer's individual CSF. Four different images at each of five different contrasts were tested. For each image, simulations were generated corresponding to views from three observation distances. For the three distances (106, 212, and 424 cm), the images spanned visual angles of 4°, 2°, and 1°, (i.e., maximum eccentricities of 0.5°, 1°, and 2°, respectively). The simulated distance and the corresponding size in degrees served to establish the proper relationship between the observer's CSF, expressed in c/deg., and the image spatial frequencies, expressed in c/image. The observers viewed the image pairs from nine distances, which included distances both shorter (53 cm) than the shortest simulated distance and longer (848 cm) than the longest simulated distance. Each image at each simulation distance

was presented 10 times at each viewing distance. The position of the simulated image relative to the original (right or left) was randomly selected for each presentation. From each observation distance, the Percent Correct identification of the processed/ simulated image was calculated for each of the four test images at each of the simulated distances. The data for each simulated distance (Percent Correct out of 40 responses for each observation distance) was fitted with a Weibull psychometric function to determine threshold at a 75% correct level. The distance at which the observers performed at the 75% level was compared to the simulated distance. If the simulations and the CSF used in the simulation accurately reflect the observers' perception, the measured and simulated distance should be equal.

The CSF data used in the simulations were obtained for each observer individually using 1-octave Gabor test stimuli. The CSFs were obtained using a VisionWorks™ system (Durham, NH) with an M21LV-65MAX monitor (DP104 phosphor) operating at 117 Hz, non-interlaced. Method-of-Adjustment (MOA) and Staircase procedures were used, as indicated. Seven interwoven frequencies, separated by one octave between 0.5 and 32 c/deg., were presented in each block. For the MOA, a threshold was estimated by averaging six responses at each frequency. For the Staircase procedure, six response reversals were obtained and a threshold was estimated from the mean of the final four reversals. The stimuli were the same 1-octave, Gabor patches of bandwidth in all cases (vertical orientation only). In a previous study [15] CSF data were obtained using both an orientation detection task and a pattern detection task. In the orientation detection task, a Gabor patch was presented in a single testing interval and the observer was asked to make a forced-choice response as to whether the grating was horizontal or vertical. The pattern detection task was performed in a temporal, two-alternative forced choice and the observer indicated which interval contained the stimulus. In both cases, a staircase procedure was used. The results reported by Peli [15] clearly rejected the simulation based on the orientation-detection CSF, and so here we report only results obtained using simulations based on the pattern-detection CSF.

The image pairs were presented on a 19 in (48 cm), non-interlaced monochrome video monitor of a Sparc 10 Workstation (Sun Microsystems, Mountain View, CA). The display luminance was linearized over a two log-unit range using an 8-bit lookup table. The images were 256 × 256 pixels each, and were presented at the middle of the screen, separated by 128 pixels. The background luminance around the images was set to the mean luminance level of the display (40 cd/m^2). The test images were also produced at varying contrasts by subtracting the mean luminance level from the image, multiplying each pixel by the corresponding contrast (10%, 30%, 50%, and 300%), and then adding the mean luminance back. The 300% contrast image was saturated wherever the dark or bright values exceeded the dynamic range of the display. The simulations for each of the four images, five contrast levels, and three simulated observation distances were generated as described in Peli [8].

Observers were seated in a dimly lit room and allowed to adapt for five minutes to the mean luminance of the display. A sequence of image pairs was then presented, and the observer responded as to the spatial location of the simulated image (right or left) by depressing the right or left button on a mouse. A new pair of images was presented 0.1 sec. after each response and remained on until the observer responded. The order in which the observation distances were tested was randomized

3.2 Results

The first set of experiments was conducted with simulated test images produced using CSF data measured from a fixed, 2 m, observation distance. The CSF values needed for the simulations at frequencies outside the measured range (0.5 - 16 c/deg.) were extrapolated by linearly extending the low and high frequencies limbs of the CSF. The CSF was measured using the MOA. For the well-trained psychophysical observers the results with the MOA and Staircase procedures differed only slightly. Both the form of the CSF and the standard error of the measurement, were similar to those of CSFs obtained using similar stimuli, similar forced choice procedures, but different display systems. This was not the case for one novice observer (JML) whose staircase data were similar to those of the other observers, but whose MOA data showed substantially reduced sensitivity (as much as 0.5 log units at middle and low frequencies), even when measured repeatedly.

Four observers participated in this experiment and their results were similar. Shown in Fig. 2 are data obtained from observer AL. If the simulations were veridical, the fitted curves would cross the 75% correct level at the simulated distance, and thus all points in Fig. 2 would lie on the diagonal line. However, as can be seen from the figure, the results the simulation was veridical only for the images in the 30 - 100% contrast range, even for the most practiced observer (AL, who participated in a pervious study using the same task). For these moderate contrast images, the distance at which the original was distinguished from the simulation was very close to the simulated distance. The 10% image was discriminated at distances larger than the simulated distances, indicating that the CSF values used in the simulations at low frequencies were too low. Stated otherwise, the thresholds implemented in the simulations were too high, in that they removed more of the image than was appropriate, thus making the discrimination task easier. The 300% image was discriminated at a shorter distance, indicating that the CSF values used at the high frequencies were too high. The results for a second experienced observer (KB), who was however a novice at this task, are generally similar except that performance was somewhat poorer in that shorter observation distances were required to distinguish the simulated image from the originals. In addition, the data for this observer differ even more at the moderate contrast levels than do those of observer AL. The results for the remaining two observers were similar to those of observer AL, in that they were centered on the diagonal prediction line. However, the variability of the data of the latter two observers was greater than that of observer AL, and was more similar to that of, observer KB.

By varying image contrast in the present study, we were able to test the CSF over a wider range of frequencies than was tested by Peli [15]. Although we do not present all of the data here, we note that the individually measured CSFs used in the present simulation under-estimated the observers' sensitivity at low spatial frequencies and

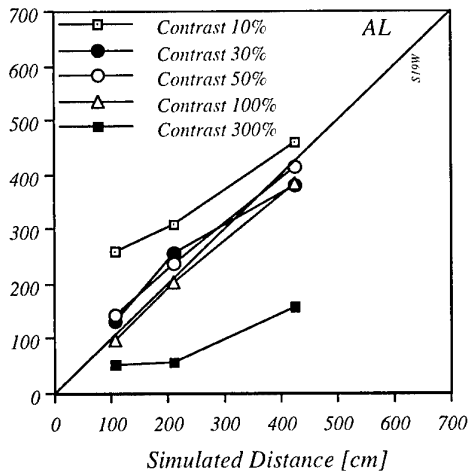


Fig. 2. The measured distances (y-axis) at which the simulated images were just distinguishable from the corresponding original images, plotted as a function of the simulated observation distances. For this observer, the data deviate from the veridical (diagonal line) only for the extreme contrast conditions—one (10% contrast) corresponding to detection of low spatial frequencies and the other (300% contrast) to high spatial frequencies. The data from the other observers were similar in form, although their variability was somewhat greater.

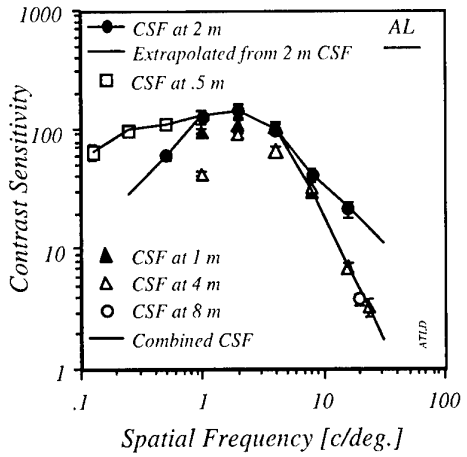


Fig. 3. Contrast sensitivity (y-axis) measured for two observers at different observation distances. The 2 m data together with the illustrated extrapolations were used in the simulations of the first experiment. The data shown with a solid line labelled "combined CSF" were used in the second experiment. The data for the second observer were essentially the same as those shown.

over-estimated it at high spatial frequencies. Since extrapolated CSF values at both ends of the frequency range were used in the simulation, further experiments were conducted to determine if the observed deviations from the expected distance estimates (i.e., the diagonal line in Fig. 2) at low and high contrasts was a result of an error in the CSF measurements, or simply an error in our extrapolation of those measurements.

The contrast sensitivity was re-measured for two of the four observers using the same stimuli, procedure, and display system, but varying the observation distance to extend the spatial frequency range. The smallest observation distance tested was 0.5 m, which reduced the lowest spatial frequency tested from 0.5 c/deg. to 0.125 c/deg. The three lowest frequencies were obtained using the 0.5 m viewing distance. The greatest observation distance tested was 8 m, which permitted testing at frequencies as high as 24 c/deg. (our observers could not detect the 32 c/deg. stimuli at any contrast). As can be seen in Fig. 3, contrast sensitivity to the lower frequencies, measured at the smallest observation distances (square symbols), were higher than the previous measurements. This result is consistent with the simulation results of Experiment 1. Also, the contrast sensitivity to the higher frequencies measured at the greater observation distances of 4 and 8 m were almost overlapping. These high-frequency data, showed substantially lower sensitivity than the data obtained or extrapolated from the 2 m measurements. These differences insensitivity are also consistent with the results obtained in the simulations, suggesting that the contrast sensitivity of the observers in this task is better represented by the directly measured CSFs than by data extrapolated from the CSF obtained at 2 m. Similar conclusions could be drawn from the data obtained for the second observer (KB). It should be noted that except for the 24 c/deg. data, the new measurements were obtained using the same physical stimuli as were used at the 2 m distance.

To further verify the simulations and to better determine the most appropriate CSF for use in simulations of this kind, we repeated the first experiment for two observers. In the second experiment we used the CSF shown in Fig. 3, which was obtained by combining the data from various observation distances. Specifically, the 0.5 m measurements were used for the low spatial frequencies, the 2 m measurements were used for the intermediate frequencies, and the 4 m measurements were used for the high frequencies. The contrast sensitivity at 32 c/deg. required by the simulation was extrapolated from values at 8, 16, and 24 c/deg., since we could not obtain contrast sensitivity measurements from our observers at that frequency. The results (Fig. 4) clearly show a convergence of the data towards the diagonal line for observer AL. Observer KB showed a substantial convergence of the data from various contrast versions and in addition a combined improvement in overall performance. This improvement may be accounted for by the increased familiarity with the task. For both observers, the deviations of the estimated distance from the predicted distance are smaller than those evident in the data of Fig. 2. In particular, the values for the 10% and 300% contrast images converge towards the other values. The results for the 300% contrast image remain separated from the rest of the samples. Since the 300% images test the CSF at high spatial frequencies, this suggests that the observers' visual performance in the task was mediated by even lower

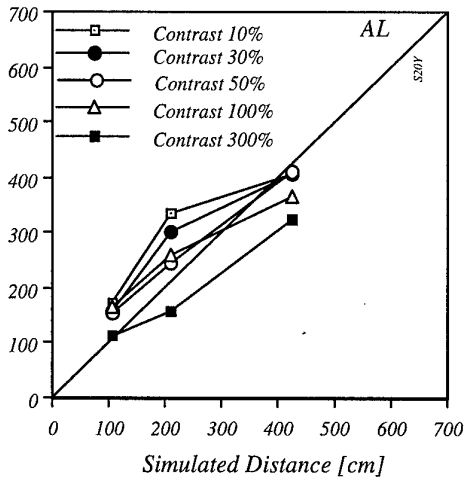


Fig. 4. The measured distances (y-axis) at which the simulated images were just distinguished from the corresponding original images compared with the simulated observation distance, for the same observers as in Fig. 2. Here the simulations were computed using the combined CSFs obtained from different observation distances (Fig. 3). For the well-practiced observer the data with the combined CSF is now very close to the prediction represented with the diagonal solid line.

sensitivities than those measured from the 4 m observation distance.

4. PERIPHERAL VISION MODEL

Non-uniform processing is a salient and well-documented feature of the visual system [17]. Peli et al. [18] showed that the changes in contrast sensitivity across the retina might play a role in maintaining size (distance) invariance i.e. they may account for the fact that "form perception is largely independent of distance" [19]. Such distance invariance has been reported for various stimuli [20-22]. The property of the visual system that allows the detection of image contrast to be nearly invariant with the changes in size associated with changes in distance must be included in any complete visual simulation.

The model we propose for describing changes in contrast sensitivity as a function of eccentricity consists of the foveal CSF and one additional parameter, the fundamental eccentricity constant (FEC). The FEC represents the decrease in contrast sensitivity as a function of retinal eccentricity [18]. Specifically, the FEC is the slope of the function relating the contrast threshold for a 1 c/deg. stimulus to retinal eccentricity, on a log-log graph. This simple relationship allows us to model the effects of visual system non-uniformity on the appearance of wide-field images using only limited data on the sensitivity of the retina at various eccentricities. We have also made use of a pyramidal, local band-limited contrast model [8], which

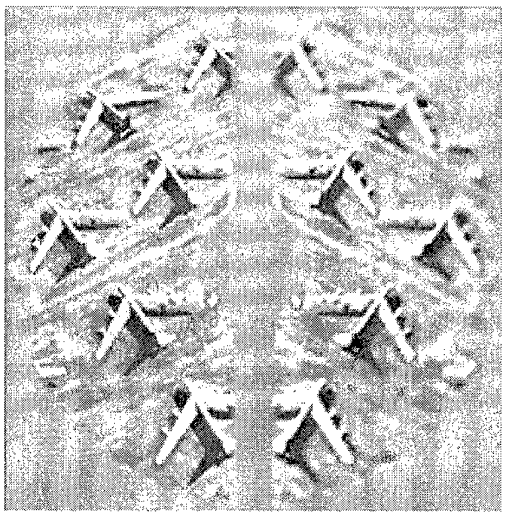


Fig. 5 Typical test stimulus. This image was obtained by applying an FEC level of 0.15 to the right side of the mirror image pair derived from the right side of the planes image.

has previously been used to model the appearance of images processed by a non-uniform visual system [18]. Here we used the discrimination of wide-field imagery to test the validity of these previously-described simulations and to determine whether a single eccentricity-dependent parameter (i.e., the FEC) is sufficient to model the well-known spatial non-uniformity of the visual system. We further attempted to determine which of several CSFs was the best estimator of the discrimination of complex, wide-field imagery.

4.1 Methods

Six observers were tested, although not all under all experimental conditions. The observers ranged in age from 18 to 48 years, and had uncorrected 20/20 vision as determined by a Snellen chart. The observers were paid for their participation.

Four stimulus images were obtained from the left and right halves of two digitized aerial photographs, one of airport buildings and the other of planes on the ground (Fig. 5). One half of each stimulus image was an unprocessed version of the original half-image, and the other was a mirror image of the original half-image (Fig. 5), processed as described below. The full stimulus images were 1024 x 1024 x 8-bits and subtended 64° at a viewing distance of 1.2 m. Stimuli were rear-projected onto a large screen (Lumiglas 130, Stewart Film Screen Corp.) using the green channel of a Barco Graphics 808s CRT. Stimulus presentation and data collection were controlled by an SGI Crimson workstation.

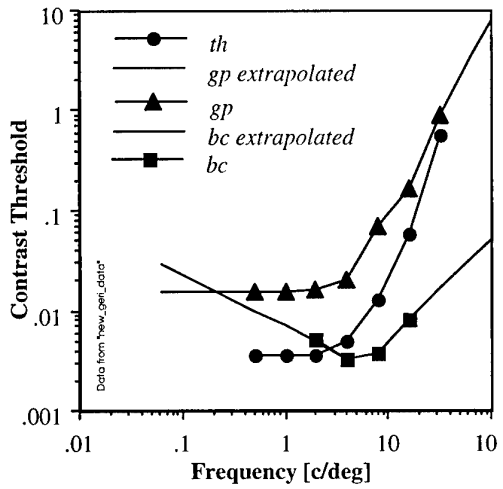


Fig. 6. The three CSF data sets used in the various simulations. The *gp (orientation detection/ variable window)* set was obtained with Gabor stimuli in an orientation discrimination task. The *th (pattern detection/ variable window)* set was obtained using the same stimuli in a contrast detection task. The *bc (pattern detection/ constant window)* set was obtained with a fixed aperture grating stimuli in a detection task. Note that the *bc* and *th* CSF are identical at mid-frequencies of about 3 c/deg. Extrapolated values were used in the simulations when needed outside the available data range.

To simulate their appearance across 64° of visual angle, the images were processed assuming fixation at their center. The details of the simulation method are given in Peli [8], and the modifications used for peripheral simulations are given in Peli et al. [18].

The appropriate threshold at each location was determined using the foveal CSF data set and the FEC applied for a given simulation. The threshold was calculated for each eccentricity, θ, and for each *FEC* as:

$$T(\theta, f) = T(0, f) \cdot \exp(FEC \cdot \theta \cdot f), \qquad (1)$$

where $T(0, f)$ is the foveal threshold and f is the spatial frequency in c/deg.

The images were processed using one of three CSF data sets (*pattern detection/ constant window, orientation detection/ variable window,* or *pattern detection/ variable window*) and one of seven FEC levels (0.02, 0.035, 0.055, 0.075, 0.10, 0.15, and 0.20). The FEC levels of 0.035 and 0.055 were found by Peli et al. [18] to fit various peripheral CSF data sets from the literature. The remaining FEC levels were selected to cover a suitable range around these values. The *orientation detection/ variable window* CSF data were based on the discrimination of horizontal and vertical 1-octave Gabor patches (i.e., a sinusoid within a gaussian aperture) and were low-pass in character. The *pattern detection/ variable window* CSF data were obtained using a contrast detection task and similar stimuli [15]. These were also low-pass in character. The *pattern*

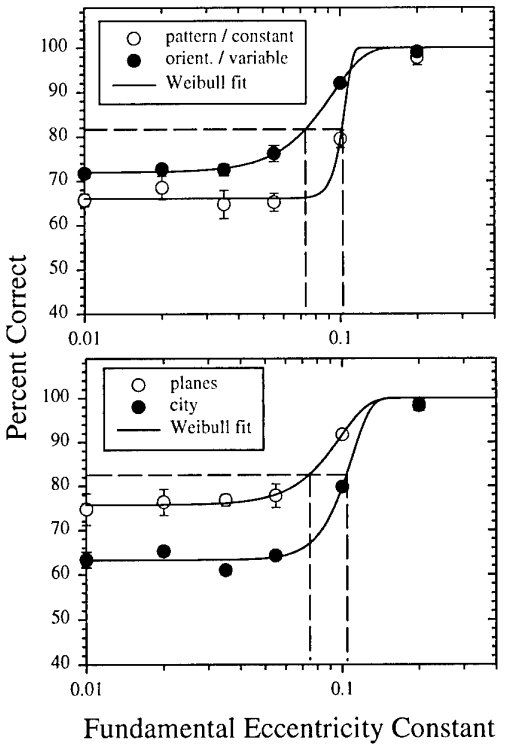


Fig. 7 Preliminary results reported by Peli and Geri [16] and Peli [24]. The data fall on the active portion of the psychometric function, and the threshold FEC is close to the prediction. However, the percent correct is high even for very low levels of FEC (upper graph), and there is a significant image dependence (lower graph). Both effects are inconsistent with the present vision model.

detection/ constant window CSF data were based on contrast detection of sinusoid gratings within a 2° square aperture [23]. These data were band-pass in character and the absolute values for the mid-spatial frequency range were similar to that of the *pattern detection/ variable window* data (see Fig. 6). Whenever values outside the measurement range were needed for the simulations they were extrapolated as shown.

A total of 560 trials were run in each 1-hr. session. The 560 trials corresponded to 10 random presentations of each of 56 stimulus images (i.e., 4 original images × 2 sides for the standard × 7 FEC levels). The data presented here are means of five Percent-Correct estimates, each in turn obtained from the forty responses within an individual session.

4.2 Results

Preliminary results of this study, shown in Fig. 7, have been reported in part by Peli and Geri [16] and Peli [12]. The smooth curves in the graphs represent the best-fitting, two-parameter Weibull function. The basic finding that the chosen FEC range resulted in a full psychometric function indicated to us that the simulations were approximately correct.

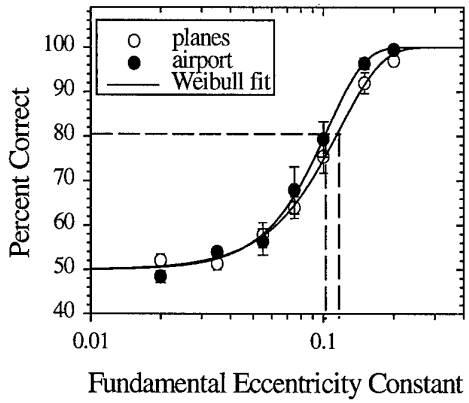


Fig. 8 Effect of the high-frequency residual on image dependence. Once the HFR (that was included in the images used for the results shown in Fig. 7) was removed, the image dependence was significantly reduced.

A surprising result was that even for very low FEC values the discrimination of the simulation from the original was at a level much higher than the 50% chance. Such a result is possible since the images are processed by the foveal CSF even for an FEC of zero, and therefore they differ from the originals. However, it is important to note that for the low values used, the simulations were processed so little that they were difficult to distinguish when examined carefully side-by-side on the screen for unlimited time. Initially we suspected that the high level of discrimination in the periphery was a result of the abrupt short presentation [25]. However, changing the presentation waveform to a 500 msec gaussian did not change the results.

Shown in Fig. 7 (lower) are means obtained from four observers for the two images used. The data indicate that the "planes" image simulation was easier to discriminate from the original than was the "city" image. Since the vision model is observer-based and includes no image-dependent parameters, it cannot account for this aspect of the data. We have seen similar effects in testing simulations of central vision [15], and in that case the effects were attributed to an artifact, the so-called high frequency residual (HFR), which was removed from the simulations but which remained in the original image. The HFR is the set of spatial frequencies at the corners of the square spatial-frequency support, which are excluded when only a circularly symmetrical filter is used. Peli [15] found that removal of the HFR resulted in the elimination of the image dependency as well as an improved performance of his simulations at various viewing distances. The data shown in Fig. 8, again are means obtained from four observers, but were obtained using stimulus images from which the HFR had been removed. Although there is a small difference between the curves at one or two FEC levels, the image dependence has been significantly reduced.

Shown in the upper graph of Fig. 9 are the functions relating the percent correct obtained in the discrimination task to FEC level for images simulated using either the *pattern detection/ constant window* (open symbols) or *orientation detection/ variable window* (closed symbols)

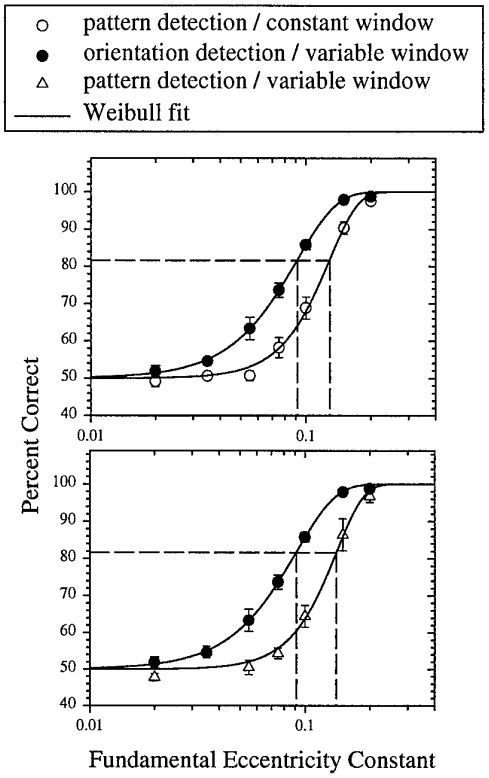


Fig. 9. A comparison of the discrimination data for images simulated using CSFs in turn obtained using various combinations of detection task (pattern or orientation) and stimulus window (constant or variable). The error bars are ±1 s.e.m. intervals about each data point.

CSF functions. These data represent averages for four observers. The FEC level corresponding to 81.6 Percent Correct was 0.128 for the *pattern detection/ constant window* data and 0.091 for the *orientation detection/ variable window* data. Analogous data comparing the results for the *pattern detection/ variable window* and *orientation detection/ variable window* CSF functions are shown in the lower graph of Fig. 9. The *pattern detection/ variable window* data were obtained for three of the four observers from whom the *pattern detection/ constant window* and *orientation detection/ variable window* data were obtained. The average threshold FEC level for the *pattern detection/ variable window* data was estimated to be 0.140.

The results in Fig. 9 show a clear difference between the simulations based on the *orientation detection* and *pattern detection* CSFs, but cannot differentiate between the two data sets based on the *constant window* and *variable window* CSFs obtained using the pattern detection task. As shown in Fig. 6, the CSFs associated with these latter two data sets converge at middle frequencies. Since the *constant window* CSF show higher threshold than the *variable window* at low frequency, which are tested by low contrast images, we can expect the *variable window* simulations results to require higher FEC values to match. Thus, we can predict that the simulations would diverge and in a predictable manner if images of lower or higher

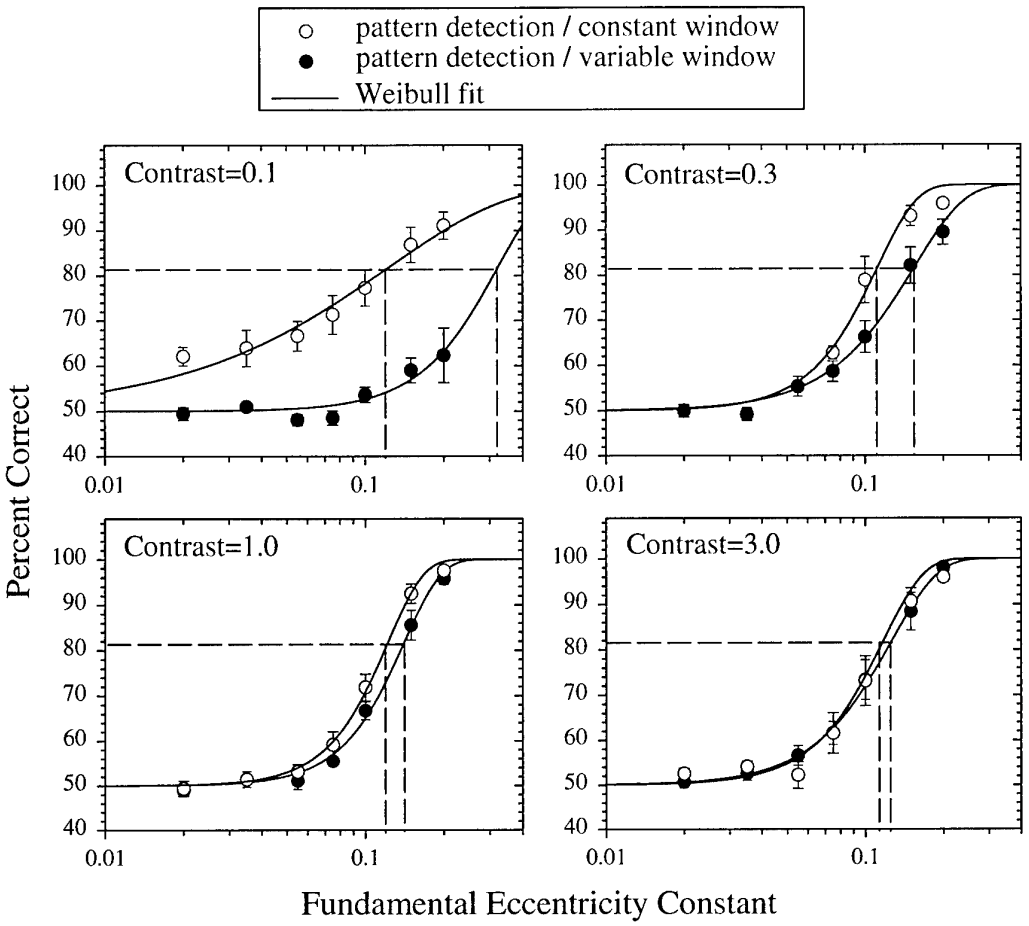


Fig. 10. The effects of image contrast on the functions relating Percent Correct discrimination to FEC level. The error bars are ±1 s.e.m. intervals about each data point.

contrast were used to test lower and higher frequencies, respectively. The one that would remain stable (if any) is the one representing the subjects' perception. The experiments therefore were repeated using test images whose contrasts (actually amplitudes) were scaled by factors of 0.1, 0.3, and 3.0 compared to the original image set. The results shown in Fig. 10 were indeed as expected. The results for the original image (contrast = 1.0) essentially replicate, using a different set of observers, the previous data shown in Fig. 9. As the contrast was reduced, the FEC found for the constant-window condition remained largely unchanged (even though the slope of the psychometric function changed, especially for contrast = 0.1). For the variable-window condition, however, the FEC gradually increased as contrast was reduced. These results suggest that peripheral sensitivity decreases at lower spatial frequencies. This conclusion is consistent with the data of Fig. 10 obtained using images simulated with the constant-window CSF (open circles), and is not consistent with the data obtained using images simulated with the variable-window CSF (filled circles). In all cases, the FEC found in our experiments was higher than the 0.035 - 0.055 value we computed from a number of data sets published in the literature, where grating targets on a uniform background were used [18].

5. DISCUSSION

The results of the foveal experiments verified that the model proposed by Peli [8] and used to simulate the appearance of an image from different observation distances is valid. The simulated images were found to be distinguishable from the original images at distances close to the simulated distances. Thus, these results demonstrate that the present simulation procedures can be used to determine whether the discrimination of complex images can be predicted from the form of empirically derived CSFs.

Also of interest are the possible reasons for the differences between the CSF's obtained here for different observation distances. Differences at the low frequency end are relatively easy to account for. The low frequency Gabor stimuli viewed from a distance of 2 m are quite large, and often extended to near the edge of the display area. The edge of the screen (outside the display areas) is dark and thus creates a high contrast feature which when close to the stimuli may reduce their visibility. Moving the observer closer to the screen reduces the size of the stimuli providing the same spatial frequencies, thus increasing their distance from the display edge and reduces its masking effect. Indeed for both observers the detection threshold for the three lowest spatial frequencies was almost equal at 2 m and 0.5 m (which were the same physical stimuli) suggesting that the low-frequency

reduction in sensitivity to these stimuli is, in fact, a masking effect. This argues for an even higher sensitivity at low frequencies than that represented by the "combined CSF" in Fig. 3. The reduction in high-frequency sensitivity that we found when observation distance was increased cannot be as easily explained.

The larger stimuli viewed at 4 or 8 m were closer to the edge, and hence may have been masked as the stimuli used to test the lower frequencies were assumed to be. These images were also of higher contrast, which might be hypothesized to result in more masking. However, such masking should reduce the sensitivity to these frequencies - not improve it. Thus, at the moment we have no hypothesis that adequately explains this effect.

We found that by varying a single parameter (the FEC) in our simulation, we could vary discrimination gradually from chance level to 100%. This confirms the validity of the simulation technique and suggests that the FEC adequately represented the position-dependent changes in the appearance of our stimuli. The fact that wide-field images simulated using various foveal CSFs were discriminated at different FEC levels, extends to peripheral vision the findings of Peli [15], regarding the sensitivity of both the simulation and the psychophysical method to small differences in the CSF.

The CSFs obtained with pattern detection tasks resulted in lower thresholds, thus it is not surprising that the simulations using the *pattern detection/ constant window* and *pattern detection/ variable window* CSFs led to higher FEC levels. The combination of lower foveal thresholds and higher FECs compensate for each other in computing peripheral thresholds. The foveal CSFs used in the simulations were complete functions of contrast sensitivity in that they differed in shape as well as absolute sensitivity. One may wonder how is it possible that such large differences in the form of the CSF can be compensated by changes in a single variable, the FEC, such that the threshold of discrimination remains the same. This is the case because, as pointed out by Peli [15] and confirmed by Peli [26], while the whole CSF is used in the generation of the simulation, only a very narrow range of frequencies (probably near the contrast sensitivity peak) is tested in the image discrimination task. The finding that the *pattern detection/ variable window* CSF, which is very similar to the *pattern detection/ constant window* CSF for mid-frequencies (Fig. 6), resulted in similar threshold FECs, also further supports this conclusion.

All FECs found here are larger than those computed from CSF measurements obtained across the retina (see Table 1 in Peli et al. [11]). This difference may be due to masking by superimposed or adjacent image detail, which may result in a more rapid decline of detection performance as retinal eccentricity is increased (i.e., crowding or lateral masking effects). While such effects were demonstrated in a number of letter acuity tests, we are not aware of direct measurements of such lateral masking for gratings or grating patches.

The vision model employed here produced images that were discriminable in a way predictable from well-established visual and perceptual data. It might, therefore, be expected that analogous models could be employed to assess more general image properties such as perceived image quality. In particular, the peripheral vision model suggested by the present analysis might be useful in evaluating wide field simulator images as well as area of interest, or other foveating systems.

ACKNOWLEDGMENTS

This work was supported in part by grants EY05957 and EY10285 from the NIH, and by U.S. Air Force Contract F41624-97-D-5000 to Raytheon Training and Services Co. at the Air Force Research Laboratory, Mesa, Arizona. We thank Brian Sperry, Jack Nye, and Craig Vrana for programming support.

REFERENCES

1. Pelli, D., *What is Low Vision?* 1990, Institute for Sensory Research, Syracuse University: Syracuse, NY.
2. Lundh, B.L., Derefeldt, G., Nyberg, S., and Lennerstrand, G., "Picture simulation of contrast sensitivity in organic and functional amblyopia", *Acta Ophthalmologica (Kbh)*, 59, pp. 774-783, 1981.
3. Ginsburg, A.P., *Visual information processing based on spatial filters constrained by biological data,* in *Aerospace Med. Res. Lab. Rep.* 1978, Cambridge University: Wright-Patterson AFB, OH.
4. Thibos, L.N. and Bradley, A., "The limits of performance in central and peripheral vision", *Digest of Technical Papers Society for Information Display*, 22, pp. 301-303, 1991.
5. Larimer, J., "Designing tomorrow's displays", *NASA Technical Briefs*, 17(4), pp. 14-16, 1993.
6. Lubin, J., "A visual discrimination model for imaging system design and evaluation", in *Vision Models for Target Detection,* E. Peli, Ed. pp. 245-283, World Scientific, Singapore, 1995.
7. Peli, E., Goldstein, R.B., Young, G.M., Trempe, C.L., and Buzney, S.M., "Image enhancement for the visually impaired: Simulations and experimental results", *Investigative Ophthalmology and Visual Science*, 32, pp. 2337-2350, 1991.
8. Peli, E., "Contrast in complex images", *Journal of the Optical Society of America [A]*, 7, pp. 2030-2040, 1990.
9. Duval-Destin, M., "A spatio-temporal complete description of contrast". in *Digest of Technical Papers Society for Information Display*, Society for Information Display, 1991.
10. Daly, S., "The visual differences predictor: An algorithm for the assessment of image fidelity", *Proceedings of the SPIE Vol. 1666 Human Vision, Visual Processing, and Digital Display III*,, pp. 2-15, 1992.
11. Peli, E., Yang, J., and Goldstein, R., "Image invariance with changes in size: The role of peripheral contrast thresholds", *Journal of the Optical Society of America [A]*, 8, pp. 1762-1774, 1991.
12. Peli, E., "Simulating normal and low vision", in *Visual Models for Target Detection and Recognition,* E.Peli, Ed. pp. 63-87, World Scientific Publishers, Singapore, 1995.

13. Stephens, B.R. and Banks, M.S., "The development of contrast constancy", *Journal of Experimental and Child Psychology*, 40, pp. 528-547, 1985.

14. Brady, N. and Field, D.J., "What's constant in contrast constancy? The effects of scaling on the perceived contrast of bandpass patterns", *Vision Research*, 35(6), pp. 739-756, 1995.

15. Peli, E., "Test of a model of foveal vision by using simulations", *Journal of the Optical Society of America A*, 13, pp. 1131-1138, 1996.

16. Peli, E. and Geri, G., "Putting simulations of peripheral vision to the test", *Investigative Ophthalmology and Visual Science.*, 34 (4, suppl), pp. 820, 1993.

17. Schwartz, E.L., "Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception", *Biology and Cybernetics.*, 25, pp. 181-194, 1977.

18. Peli, E., Yang, J., Goldstein, R., and Reeves, A., "Effect of luminance on suprathreshold contrast perception", *Journal of the Optical Society of America [A]*, 8, pp. 1352-1359, 1991.

19. Fiorentini, A.L., Maffei, L., and Sandini, G., "The role of high spatial frequencies in face perception", *Perception*, 12, pp. 195-201, 1983. (esp. p. 196)

20. Hayes, T., Morrone, M.C., and Burr, D.C., "Recognition of positive and negative bandpass-filtered images", *Perception*, 15, pp. 595-602, 1986.

21. Norman, J. and Ehrlich, S., "Spatial frequency filtering and target identification", *Vision Research*, 27, pp. 87-96, 1987.

22. Parish, D.H. and Sperling, G., "Object spatial frequencies, retinal spatial frequencies, noise and the efficiency of letter discrimination", *Vision Research*, 31, pp. 1399-1415, 1991.

23. Cannon, M.W., Jr., "Perceived contrast in the fovea and periphery", *Journal of the Optical Society of America [A]*, 2, pp. 1760-1768, 1985.

24. Peli, E., *Visual Models for Target Detection and Recognition*, Series on Information Display, Ed. H.L. Ong, Vol. 1. World Scientific Publishers, Singapore, 1995.

25. Peli, E., Arend, L., Young, G., and Goldstein, R., "Contrast sensitivity to patch stimuli: Effects of spatial bandwidth and temporal presentation", *Spatial Vision*, 7, pp. 1-14, 1993.

26. Peli, E., "The contrast sensitivity function (CSF) and image discrimination", in *SPIE Proceedings Human Vision and Electronic.* In press, 1999.

MODELLING OF TARGET ACQUISITION WITHIN COMBAT SIMULATION AND WARGAMES

Jan Vink
TNO-FEL
P.O. Box 96864
2509 JG The Hague The Netherlands
Tel: +31 70 3740126
Fax: +31 70 3740642
E-mail: J.K.Vink@fel.tno.nl

1. SUMMARY

This paper describes the target acquisition process from the perspective of modelling target acquisition as a part of modelling combat.

Exchanging fire obviously is very important in combat. Conditions for direct fire are line-of-sight (LOS) and some kind of perception of the intended target. LOS is deterministic and can be calculated if there is a good digital representation of the terrain. But perception is considered a stochastic process with probabilities depending on the current situation. In most stochastic combat simulation programmes and wargames there is a module that models detection and perception.

Because of the dynamic character of combat situations for observing are changing rapidly. The models are calculating situations every x seconds (typical 5 - 30). Within such a timeframe occurrences of events and the effects of these events are calculated. Illustrative events are new observations, firings, etc.

The target acquisition module is responsible for an actual list of observations. Each time-frame the list is updated: old observations are checked (observers or targets can be killed or moved) and new observations can be added. Because of the dynamic character only calculations are made for the coming time-frame. For each observer and each potential target an observation probability is calculated and comparison with a random number determines if the considered observer/target will lead to a new observation. Input for this module are elements of the situation at hand and characteristics of observer (such as the sensor used) and target (such as its dimensions).

This paper addresses some of the limitations and problems of the current implementation of the target acquisition module.

Keywords: combat simulation, wargame, detection modelling

2. INTRODUCTION

Simulating combat is a way of gaining quantitative insight in combat. Analysing historical data is another way of gaining insight, but this paper will not address that method.

Combat simulation models can be classified in different ways. One classification is the level of the military unit/system that can be simulated in the model. This can vary from the simulation of one system (eg. a sensor and its target) up to theatre level (the Gulf War). The examples used in this paper are at division and battalion level. These examples were chosen because both models have separate modules for detection. In higher level models detection is a very abstract process (how can detection between two brigades be defined?).

This paper will shortly describe methods of combat simulation and then two examples will be described, first in general and then the detection module in more detail. The paper ends with some general requirements we have for a detection module.

In this paper detection is used as a general term. It can be defined as the level of perception needed for follow-on actions.

3. SIMULATING COMBAT

Combat is a very complex system and is impossible to predict. What can be done is gaining some quantitative insight which can support decision making. Simulation is a way of gaining this insight. Big advantage of simulation is the control over the conditions, the possibility to simulate non existing systems and unbiased interpretation of the outcome.

Simulating combat, why:
- (Weapon)system procurement
- Training & education
- Doctrine development
- Evaluating possible courses of action
- Force structure

Simulating combat, how:
- Field exercises
- Wargames
- Computerised combat simulation models
- Analytical models

The methods differ in detail of representation, the influence of the man in the loop (repeatability) and speed.

Computermodels are used for wargames, combat simulation and analytical models. This paper will go into more detail in the detection modelling in some of the models we, at TNO-FEL, are working with. Please keep in mind that the detection module is only one of several modules the models are built upon.

4. EXAMPLES

4.1 The wargame Kibowi:

KIBOWI is a wargame, which is in use in the Dutch Army to support training & education of military staff. In a Kiwowi assisted exercise the military staff is out in the field, just like they would be in reality. They use their normal command & control equipment and procedures to prepare and give orders.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

These orders go to the so-called lower-control. The lower control is organised in different cells (up to 50). Each cell has the control over a number of computer-represented units. The lower-control can give orders to these units and can get information about these units. This enables the lower-control to execute the orders from the staff. On the computer the lower-control has only access to the units under control.

Every timestep (typical 10 second) the computer evaluates all the orders and calculates line-of-sight, detections, firings, losses, stocks, etc.

If the computer determines there is detection, the detected units will pop-up on the relevant control-station. The lower-control concerned can report this detection to the staff, so they will be able to respond on the current situation. Another possibility for the lower-control is defining and sending a indirect fire-request to the cell that controls the artillery units.

Because of the performance-requirements of this wargame, complex calculations must be as limited as possible. There can be thousands of units and that can make millions of possible interactions. The terrain is represented as a 100*100m grid with a height and a terrain feature for each grid. Line-of-sight calculations are done separately and take a lot of time, so calculating detections cannot be too complex.

4.1.1 Detection calculation:

For every possible interaction two distances are calculated: a minimum- and a maximum distance (d_{min} and d_{max}). Detection probability is a function of these distances:

- If the distance d between the observer and the target is less or equal as d_{min}, then the detection probability = 1.
- If $d > d_{max}$ then the detection probability = 0
- Between the minimum and maximum distance the detection probability decreases linearly

The detection probability is compared with a drawing from a uniform distribution and this determines whether there is detection or not. The two distances depend on different factors:

- The default values for this observer-target combination (distances in an open area on a clear day)
- The status of the observer (moving, mounting, in assembly area, etc.)
- The status of the target (moving, in prepared position, etc.)
- The terrain feature around the target
- The weather
- The time of the day

Advantages of this method:

- Speed: all factors are in lookup-tables and almost no calculations are needed
- Simplicity: method is very easy to understand

Disadvantages of this method:

- Is the linear relation maybe oversimplified
- No interdependencies between factors
- No false alarms and no misperception

4.2 The combat simulation model FSM (Force Structure Model):

The most important characteristics of FSM are:

- Stochastic
- Closed (no man in the loop during simulation)
- Using 100*100m grid with height information and terrain feature on each grid
- Combination of timestep and event-driven. Typical timestep is 30 seconds.

One of the input-files is a scenario description. This file is written in a special designed language and contains orders and conditions for each unit. Every time-step orders and conditions are evaluated, for each unit new positions are calculated and new statuses are determined. After that the consequences of direct and indirect fire in that timeframe are calculated. For each shooter/target combination (called monel: one-sided duel) the next conditions are checked:

- is there line-of-sight between shooter and target
- is there a detection from shooter to target
- is the shooter allowed to fire
- is the target within the range of the weapon
- does the shooter has ammunition to shoot on this target

In case a shooter has more than one target, there are some decision-rules in the model to select which one to shoot at first. If a monel is selected an event is created. This event is scheduled according the time of arrival of the ammunition. The time of arrival is defined as the moment of detection + aiming-time + the flight-time. The list of events is evaluated in time-order and scheduled events can be deleted from the list in case of a firepower kill of a shooter before time of departure of its shot. After an event is handled a new event can be created. This can be a next shot against the same target or a shot against a new target. As long as the timeframe is not over, events are handled and created. If the timeframe is over, the computer starts with the next one.

The moment of detection is very important in this process. This means that the detection module not only calculates a probability of detection, it also determines a moment of detection. We use the following formula:

$$\text{Prob}(t_{det} \leq t) = 1 - e^{-t/T_{mean}}$$

T_{mean} is dependent from a lot of factors: distance, intrinsic contrast, weather, detection aids, statuses of observer and target, etc. With this formula the moment of detection is drawn randomly. If this moment is within the timeframe concerned then there is a detection. If the moment is not in the timeframe it will be ignored.

One of the reasons using this formula is that the length of the timeframe does not influence the probability function of the detection moment. A drawback is that the probability of detecting a target (p_t) ever goes to 1.

The way T_{mean} is calculated is based on a report of Night Vision Laboratories: "Simulating combat under degraded visibility". It calculates the number of resolvable cycles across the target and uses this to determine a probability of detecting the target within one glimpse. Next p_t and a mean time under the condition of detecting are derived. T_{mean} is calibrated with the last two variables.

Known shortfalls in the current implementation:

- The only memory implemented is that detections are not lost if LOS is not interrupted. If LOS is interrupted the detection will be forgotten
- The intrinsic contrast does not depend on the position in the field

- A group of vehicles is regarded as a group of independent targets
- There are no false alarms
- The number of resolvable cycles required for 50% detection is constant in different situations (increasing this number means that for example recognition is needed for firing)
- There are no mistakes. Mistakes could lead to using the wrong ammunition on the wrong target.
- Shooter and observer are at the same position

5. REQUIREMENTS FOR THE DETECTION MODULE

We are always standing open for implementing a better detection module. There are a few basic requirements the module must meet:

- Generic, a sensor has to be specified by a few characteristics, not by a model
- Simple, the level of detail must be in line with the level of detail in other modules
- Fast, the advances in computer-power will be used for simulating more complex situations as well as for a more detailed implementation
- The module must be able to cope with dynamic situations. Units move around, fire, etc
- The detection probability may not depend on the size of the timestep
- The detection module must be in line with the other modules

THE DEPLOYMENT OF VISUAL ATTENTION: TWO SURPRISES

Jeremy M Wolfe
Brigham and Women's Hospital and Harvard Medical School
221 Longwood Ave., Boston, MA 02115 USA
wolfe@search.bwh.harvard.edu

1. SUMMARY

The visual system is not capable of processing of all aspects of a scene in parallel. While some visual information can be extracted from all locations at once, other processes, including object recognition, are severely limited in their capacity. Selective attention is used to limit the operation of these limited-capacity processes to one (or, perhaps, a few) objects at a time. Searching for a target in a scene, therefore, requires deployment of attention from one candidate target to the next until the target is found or the search is abandoned. Common-sense suggests that distractor objects that have been rejected as targets are marked in some fashion to prevent redeployment of attention to non-target items. Introspection suggests that sustained attention to a scene builds up a perception of that scene in which more and more objects are simultaneously recognized.

Neither common-sense nor introspection are correct in this case. Evidence suggests that covert attention is deployed at random among candidate targets without regard to the prior history of the search. Rejected distractors are not marked during a search. Prior to the arrival of attention, visual features are loosely bundled into objects. Attention is required to bind features into a recognizable object. For an object to be recognized, there must be a link between a visual representation and a representation in memory. Our data suggest that only one such link can be maintained at one moment in time. Hence, counter to introspection, only one object is recognized at one time. These surprising limits on our abilities may be based on a trade off speed for apparent efficiency.

Keywords: Vision, visual attention, visual search, guided search, memory, object recognition, human experimental psychology

2. INTRODUCTION

Faced with a new scene, we immediately see something. However, we do not immediately perceive everything. Thus, you might emerge from customs at the airport to be faced with a crowd of faces, one of whom should be the friend who has come to pick you up. It is not possible to simultaneously process all of the faces (not to mention the other objects in the scene) to the point of recognition. As a result, you need to search. Search from face to face in an apparently serial manner (29; 38) will either lead you to your friend or will lead you to the bus and to a reassessment of the nature of friendship.

Two aspects of the course and consequence of such a search are the topics of this paper. First, it seems reasonable to assume that, if you deploy attention to a face and determine that it is not your friend, that you will somehow mark that face so as to avoid revisiting it. Second, even if you do not recognize multiple objects when first confronted with a new scene, it seems intuitively clear that, after prolonged search, the visual scene will contain multiple, simultaneously recognized objects. The purpose of this paper is to demonstrate that neither of these reasonable hypotheses is actually true. In a field of items that are equivalent in their ability to attract attention, attention appears to be deployed at random with no regard to the prior history of deployments. When attention is deployed to an item, it becomes possible to recognize that item. However, when attention is redeployed away from the item, the item is no longer actively recognized. It may be remembered, just as an item that is out of sight is remembered. But our data indicate that simultaneous recognition of multiple objects does not occur.

This paper is organized into four sections. In the first section, we review some of the basics of laboratory visual search experiments. Next, we discuss the evidence that the deployment of attention is more anarchic than commonsense would predict. A third section considers the visual consequences of attention. Finally, the implications of these results will be discussed.

3. VISUAL SEARCH IN THE LABORATORY

3.1. Introduction to Search Methods

In a standard laboratory visual search experiment, observers search for a target item among a number of distractor items. In a typical version, the target would be present on 50% of the trials. The total number of items (the "set size") would be varied. The dependent measures are the "reaction time"(RT) - the amount of time required to press a key to indicate the

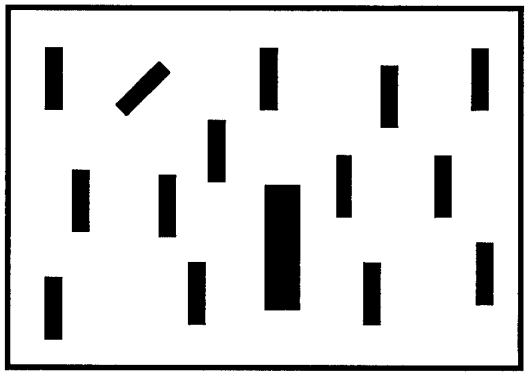


Figure One: Highly efficient search. Targets defined by salient basic features can be found, independent of the number of distractors. Here targets are defined by size and orientation.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

presence or absence of a target - and the accuracy of that response. Most of the results presented here will be RT data. The measure of greatest interest is the slope of the function relating RT to set size. This is a measure of the efficiency of search. The most efficient searches have slopes near zero, suggesting that all items can be processed at the same time, without capacity limits. Examples are shown in Figure 1.

The most efficient searches are searches for targets defined by a basic feature among homogeneous distractors (e.g. red among green, big among small, etc.) The set of basic features for visual search contains obvious candidates like color (e.g. 2; 10), size(e.g 4), and orientation (e.g 3; 26). It also contains less obvious features like lustre (61) and a variety of depth cues (14). The full list contains perhaps a dozen features (reviewed in 57).

The presence of an attribute is easier to detect than its absence. This leads to so-called "search asymmetries" (50) where the search for A among B produces a steeper slope than a search for B among A. An example is shown in Figure 2.

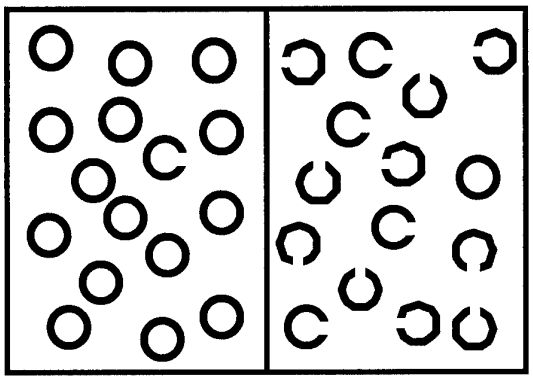


Figure Two: It is easier to find the presence of a feature (here the line terminators in the "C") than it is to find the absence of a feature. (After Treisman).

At the other end of the continuum of search tasks are inefficient searches. With slopes of about 20-30 msec/item on target present trials and about twice that on target absent trials. It is, of course, possible to have search tasks with arbitrarily steep slopes. One source of steep slopes is a need to fixate items. If the items cannot be classified as distractor or target without fixating each item, then search slopes will come to reflect the speed of eye movements (2-4 fixations per second) and will yield slopes of greater than 100 msec/ item. In experiments, like those described here, that are concerned with the covert deployment of attention, care must be taken to assure that eye movements are not required. It is less important, in most cases, to require rigorous fixation since the pattern of RTs appears to be essentially the same whether eye movements are permitted or not (66).

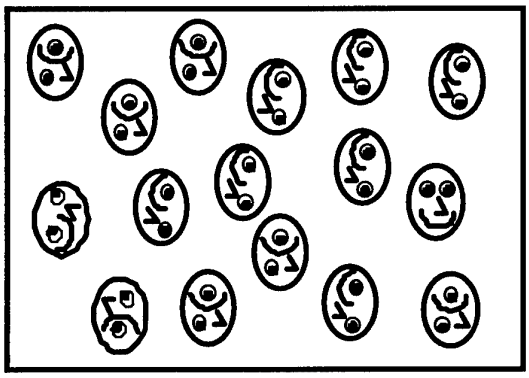


Figure Three: Even ecologically significant stimuli like faces produce inefficient search if they do not differ from distractors in basic features.

The class of inefficient searches includes all those for which basic feature information is of no use. This includes searches for easily identifiable objects like faces and animals where identification is based on the relationship of features to one another rather than to the mere presence of a defining feature. Our data indicate that the shape of an object is not a basic feature for visual search. If local features like line termination are controlled, search for one shape among other, quite different shapes is inefficient (59).

3.2. Conjunctions and Guided Search

Most natural searches are neither feature searches nor random searches among preattentively equivalent items. Most searches involve targets that, while they are not defined by a single unique feature, are defined, at least in part, by basic feature information. Thus, the hunt for your friend at the airport requires a search but it is a search through a subset of visible objects. Little time will be spent examining suitcases and car rental signs (13).

Laboratory search experiments have concentrated on the less natural case of *conjunction search*. In a typical conjunction search, targets are defined by the presence of two features (e.g. a *black vertical* target) among a mix of distractors that have one or the other of these features (e.g. white *vertical* and *black* horizontal distractors).

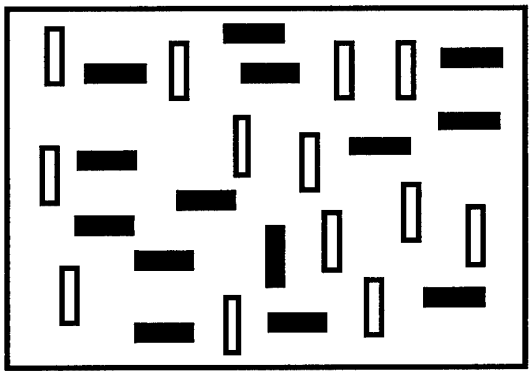


Figure Four: Conjunction search. Find the black vertical item.

Work in the 1970's and early 80's seemed to show that conjunction searches were uniformly inefficient (48). These and other data led to Treisman's very influential proposal that searches could be divided into two categories: Feature searches that could be performed in parallel and all other searches that required serial, item by item, inefficient search. This hypothesis was one of the central propositions of Treisman's original formulation of her "Feature Integration Theory" (48). However, subsequent research revealed that conjunction search could be quite efficient (e.g. 9; 24; 28; 34; 49; 60; 67). At first, it appeared that these efficient conjunctions searches might represent specific exceptions to the general rule of inefficient conjunction search (23; 27).

However, it has become increasingly clear that search for any conjunction of basic features can be efficient if the features are salient enough (see discussions in 56; 58). Indeed, there are several published reports of conjunction searches that yield search efficiencies that are indistinguishable from those produced by basic features (e.g. 40; 52; 55).

In retrospect, this is not a surprise. As the earlier example should have made clear, it is intuitively obvious that attention is somehow guided to likely targets. The Guided Search model makes the claim that this guidance comes from preattentive feature information (6; 56; 60; 62). That is, Guided Search holds that no preattentive process has explicit information about conjunctions. However, to continue with the example from Figure Four, a color processor can guide attention toward black items while an orientation processor can guide attention toward vertical items. The combination of these sources of guidance will tend to guide attention toward items that are both black and vertical (see Figure Five).

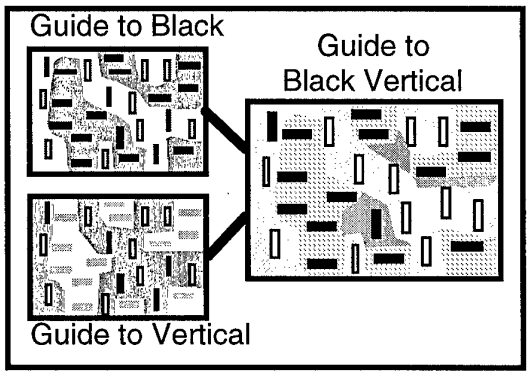


Figure Five: The core idea of Guided Search is that basic feature information can be used to guide attention to targets defined by more than one feature.

Revisions of Feature Integration Theory incorporate feature guidance (46; 49) as do some other models (e.g. 51). On the other hand, there are models, notable Duncan and Humphreys' (11) Similarity model that propose explicit preattentive processing of conjunctions.

3.3. The Myth of Two Classes of Search Tasks

The influence of the 1980 version of Feature Integration Theory has been long and wide. An unintended consequence has been the wide-spread assumption that there are two types of visual search, "serial" and "parallel" and that specific tasks can be placed in one of these two categories on the basis of the slope of the RT x set size function.

In fact, as should be clear from the preceding discussion, search tasks yield a continuum of slopes from efficient to inefficient with no value dividing these slopes into two principled groups. To illustrate this point, we pooled 2000+ search slopes from a range of different feature, conjunction, and letter searches. The distribution of slopes is shown in Figure Six.

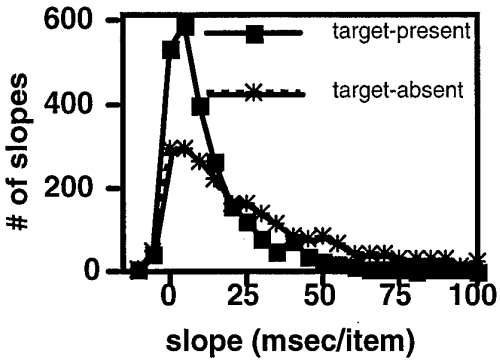


Figure Six: The distribution of 2000+ search slopes showing that there is no obvious division of tasks into search classes on the basis of slope alone (redrawn from 58)

The purpose of this exercise is not to argue that all search tasks are drawn from the same distribution. If we sort the slopes by the type of search task, it is clear that different types of task produce different distributions of slopes. Figure Seven shows the target present slopes of Figure Six broken into three broad classes of search: feature searches, conjunction searches, and searches such as a search for a "T" among "L"s that have traditionally served as benchmark "serial" tasks.

The distributions are clearly different. Thus, search slope can be predicted (albeit imprecisely) from a knowledge of the search task. It is the reverse that does not work. It is not possible to place a dividing line at, say, 10 msec/item and declare searches on one side to be qualitatively different from searches on the other.

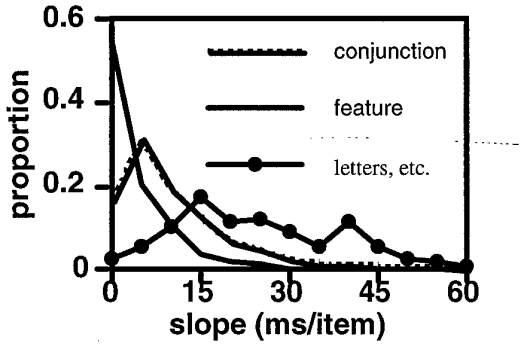


Figure Seven: Distribution of target-present slopes divided by type of task. (redrawn from 58)

There are a number of ways to understand this continuum of search slopes. In the context of the Guided Search model, all searches involve preattentive guidance of the deployment of spatial attention. For the tasks described here, the prime source of variation lies in the effectiveness of that guidance. In the most efficient feature searches, guidance is sufficient to direct attention to the target before it is deployed to any distractors. In an inefficient search such as a search for a T among Ls, guidance still limits search to the Ts and Ls. Attention is not directed to blank space or away from the search display. However, within that set of letters, there is no further guidance and search proceeds at random. Conjunction

tasks represent an intermediate case in which preattentive feature information guides attention but guides it imperfectly so that some distractors attract attention and search slopes are intermediate. In this framework, it is important to understand the rules for deployment of attention. That topic is addressed in the next section.

4. THE DEPLOYMENT OF ATTENTION: THE FIRST SURPRISE

4.1. The standard models

There are two broad classes of models of the deployment of attention. The preceding discussion has assumed a *serial* model in which attention is deployed from item to item. Alternatively, a limited-capacity resource could be allocated to multiple items *in parallel.* Guided Search generally assumes a serial model. However, in principle, preattentive processing could guide the allocation of a distributed resource rather than guiding the deployment of an item-sized attentional "spotlight". Both classes of model can predict the patterns of RTs seen in search experiments(43-45). Intermediate positions are possible. Several models propose a serial deployment of attention, not from item to item, but from one group of items to the next (e.g. 15; 31). In fact, the dichotomy between serial and parallel models may have been overstated. Consider a conveyor belt. Items may be loaded on and off the belt in series but multiple items are on the belt in parallel (see also 16; for a more extensive discussion of this idea see 25).

A hallmark of virtually all of these models of attentional deployment has been the assumption that information accumulates during the course of a trial. In serial models, this takes the form of the assumption that rejected distractors are inhibited or marked in some way so that attention is not re-deployed to previously rejected items (e.g. 1; 20; 42). Phenomena like inhibition of return (IOR) have been invoked as plausible mechanisms of distractor marking (32; 33) though efforts to find evidence for IOR in visual search have had a checkered career (18-20; 65).

In parallel models, within-trial 'memory' generally takes the form of a local accumulation of evidence over the course of a trial (in the manner of 35). Thus, in a search for a T among Ls, information about the T-ness or L-ness of each item would accumulate over time until one item was confirmed as a T or all items were confirmed as Ls. Our recent data violate the predictions of this core assumption about the deployment of attention.

4.2. The Experiments of Horowitz and Wolfe (17)

To test the hypothesis that information accumulates during the course of a visual search trial, we compared a fairly standard search with a condition designed to minimize the accumulation of information. In the first experiment, the task was a standard T among Ls search. Both Ts and Ls could appear, randomly, in any of four orientations: 0, 90, 180 and 270 deg. As usual, the subject's task was to report as quickly as possible whether or not the target letter was present in the display. Targets were present on 50% of trials. The set sizes were 8, 12, or 16. Letters subtended 1 deg at the 57 cm viewing distance.

There were two stimulus conditions in the experiments: Dynamic and Static. The Static condition was a variation on a standard visual search experiment. The stimulus presentation consisted of 20 cycles of an 83 msec presentation of the search display and a 24 msec mask composed of all of the line segments that could go into the Ts and Ls. The total stimulus duration, therefore, was 2220 msec.

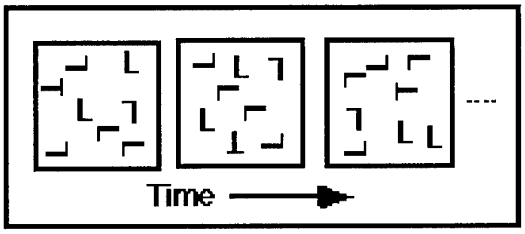


Figure Eight: The dynamic search condition. The same elements are plotted in each frame but their positions are changed randomly.

In the Dynamic conditions (shown above), the stimuli were randomly relocated every 111 msec. This did not involve any sort of coherent motion of stimuli. In this version of the experiment, a Dynamic trial consisted of five cycles of four independent frames of 83 msec duration with the 24 msec masks in between. Suppose that the trial was a target present trial with a "T" and eleven "L"s. Each of the four frames would present those twelve items in new random positions. If necessary, Ss could respond after the 2220 msec stimulus display. In practice, RTs of this length accounted for less than 2% of the data.

The Dynamic condition was intended to make any marking of rejected distractors irrelevant. If search involves serial selection of items, then the Dynamic condition should force selection *with replacement* from the set of items on the screen (That is, a given distractor might be checked more than once). The standard serial view of the Static condition has been that it involves selection *without replacement* (A given item would not be checked more than once.). In a standard serial, self-terminating search, the observer must sample an average of half of the items on target-present trials. Modeling shows that the average number of samples in the Dynamic case equals the set size. This does not mean that each item in the display is sampled. In sampling with replacement, some items may be sampled multiple times. It follows that Dynamic target present slopes should be twice as steep as the Static target present slopes, if there is marking of rejected distractors in the Static condition.

A second version of this experiment was run without the masks. In this case, the Static condition is truly static. Nine subjects were tested for 200 trials in each condition, randomly distributed over 3 set sizes.

Figure Nine shows the RT and errors as a function of set size for Experiment One. Results for Exp. 2 are comparable. The slopes for the Dynamic condition were not twice the slopes of from the static condition - falsifying the prediction of the standard serial model. Target-present slopes in static and dynamic conditions did not differ significantly in either version of the experiment. (Exp. 1: $t(8)=.13$, $p<.50$, Exp. 2: $t(8)= 1.52$, $p>.15$). Note in Figure Nine that target-absent slopes are actually *shallower* for the Dynamic case than for the Static case. While the Dynamic mean RTs do appear to be longer than the Static, that RT cost is reliable only in Experiment 2 ($F(1,8)=18.81$, $p < .005$). We suspect that the increased mean RTs reflect subjects' decreased confidence in their responses. Consider a subject who *believes* she has found a target. In the Static case, the physical stimulus is still

available for confirmation, while in the Dynamic case, it is not.

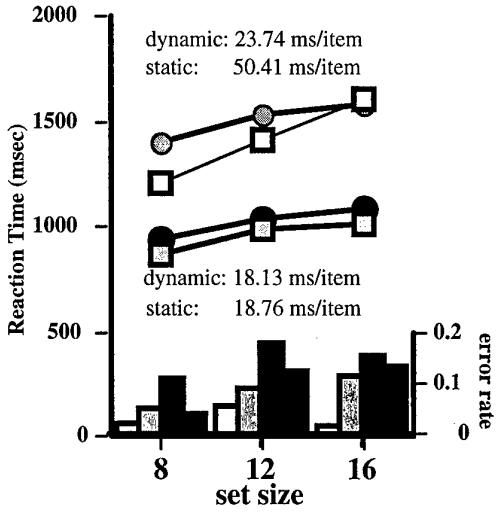


Figure Nine: Mean RT data for dynamic and static conditions of the first experiment (with masks). Upper curves are target absent. Lower are target present. Note that dynamic, target present slopes are very similar to static slopes. Bars give error rates in the following order: Static false alarms, static misses, dynamic false alarms, dynamic misses.

These results would be uninteresting if subjects, in the dynamic condition, could direct attention to one location and simply wait for the target to appear in that location. However, the position of the target was constrained in order to thwart any such strategy. In Experiment One, the target only appeared at one of four locations (one in each of the four independent frames). Here a "sit and wait" strategy would lead to failure on 93.75% of target present trials. In Experiment 2, the target changed location on every trial but remained at one of four eccentricities (again, chosen at random from trial to trial). In this case, a "sit and wait" strategy would fail on 75% of trials.

These data would have been a fairly straight-forward, if surprising, refutation of the predictions of the standard accounts of the marking of rejected distractors were it not for the error rates. Subjects make more errors in the Dynamic condition than in the Static condition. This is not surprising. Stimuli are more degraded in the Dynamic condition and, as noted in connection with the RT difference, subjects can continue to attend to a location and confirm the existence of a target in the Static condition but not in the Dynamic condition. That said, the error rates complicate the analysis of the result because of the likelihood of a speed-accuracy tradeoff. Given the more frequent errors in the Dynamic case and given the increase in those errors with set size, we must assume that the slopes in the Dynamic case are underestimates of the "true" slope. Could that "true" Dynamic slope be twice the "true" Static slope and, thus, consistent with marking of rejected distractors in the Static condition? In an effort to answer this question, we conducted a replication of the experiment with a design intended to reduce the error rates.

4.3. Experiment Three: Another Version

In this third experiment, we eliminated the option to respond "no" by having subjects respond to target identity, rather than target presence. A target letter "E" or "N" was present on each trial, embedded in distractors selected from the remaining letters of the alphabet (except for "I" and "J"). Subjects identified the target letter. Again, we compared Static and Dynamic conditions. Methods were similar to the experiments described above. Since subjects would always know that a target was present, we reasoned that they would be less inclined to abandon a difficult search with a guess. This should lower errors.

Our results showed that, once again, the slopes were statistically indistinguishable with the Dynamic slope of 29.5 ms/item being slightly shallower than the Static slope of 34.67 ms/item. The effort to reduce errors worked. Error rates were substantially lower in this experiment (5.6% overall for the Dynamic condition, 2.8% for the Static). Nevertheless, there are still twice as many errors in the Dynamic condition. Is this difference sufficient to mask a true 2:1 relationship between Dynamic and Static slopes? The point is arguable but we think that it is implausible to propose that a relatively few errors could, in effect, cut the Dynamic slope in half. It is possible, for example, to calculate the missing RTs that would be needed to double the Dynamic slope. The details of this error correction analysis are given on our website (search.bwh.harvard.edu). In brief, in order to double the Dynamic slope, one would need to assume that all errors come from trials where the reaction time should have been much longer than almost any of the correct RTs in the actual data. As a different approach, we can look at the results only for the subjects with the smallest differences between Dynamic and Static error rates. In this subset of the data, we still find that Dynamic and Static slopes are essentially the same.

4.4. Memory-free search?

How should these results be interpreted? Recall the predictions of the standard, serial, self-terminating search model. If we assume that rejected distractors are marked in the Static case and that they cannot be marked in the Dynamic case, then the target present slopes in the Dynamic case should be twice those in the Static case. The experiments yield Static and Dynamic slopes that are indistinguishable from each other. These data appear to falsify the hypothesis that rejected distractors are marked in the Static condition and not in the Dynamic condition. Given the distractors *could not* be marked in the Dynamic condition, it would seem to follow that they were not marked in the Static condition either. That is, it would appear that items are sampled from the display *with replacement* in both the Dynamic and Static cases. We have dubbed this the *memory-free search* hypothesis.

The memory-free hypothesis only applies to covert deployments of attention and not, for example, to overt eye movements. It is possible that previously fixated locations are marked in visual search (19). Covert attention and overt eye movements are usually linked (e.g. 21). Attention can be deployed at a faster rate than can the eyes. Nevertheless, some memory for prior fixation might be all the memory needed in real-world visual search. It is also important to note that the memory-free hypothesis proposes a lack of memory for rejected distractors. It does not propose a lack of memory for accepted targets. Targets must be remembered, once they are found, otherwise it would be impossible to perform repeated

searches through the same display (e.g. Where are those two kids of mine?). The act of rejecting a distractor is different than the act of accepting a target. Perhaps it is the act of coding targets into memory that produces the attentional blink (8; 36; 37).

4.5. Examining the effect of trial length

Beyond simple speed-accuracy trade-offs discussed above, there is another way for Dynamic and Static conditions to produces the same target present slopes even if search is memory-free in the Dynamic and memory-based in the Static condition. Alex Backer (personal communication) noted that the theoretical distribution of RTs is uniform and finite in the memory-based case while it has an exponential, potentially infinite upper tale. That is, suppose that a display contains ten items. In an accurate memory-based search, the observer never searches through more than ten items. In an memory-free search, however, the subject could search forever. Very long searches will be very rare, but they should occur in theory.

In practice, long RTs are less likely. After a certain point, observers will tend to give up and guess. Of more specific relevance to these experiments, Backer noted that we used 20 frames of 100 msec each. If subjects did not find a target during the 2000 msec of stimulus exposure, they would have to guess. As a consequence, RTs that would have been significantly longer than 2000 msec would have been removed from the RT distribution. Under one set of assumptions, it happens that the loss of these long RTs would be enough to reduce the theoretical slope of a memory-free Dynamic search to the slope of a hypothetical, memory-based Static search.

More generally, Backer's analysis predicts that slopes in the Dynamic condition should be strongly influenced by the duration of the stimulus display. Slopes in the Static condition are only influenced at short display durations. As a consequence, this analysis predicts that Dynamic search slopes will be shallower than Static at short durations and longer at long durations with a fairly narrow range of durations producing roughly equal slopes in the two conditions.

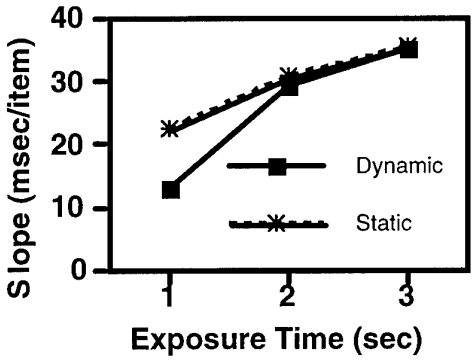


Figure Ten: Slope as a function of exposure duration of Dynamic and Static search displays. Note that the slopes converge as the duration gets longer.

In order to assess the possibility that we had inadvertently stumbled on the point of equality, we tested subjects at display durations of 1, 2, and 3 seconds. The task was the "E or N?" task described above. Methods were similar to those described for that experiment.

Figure Ten shows the results of this experiment. At the shortest duration, in partial support of Backer's hypothesis, the slopes for the Dynamic case are somewhat shallower than the slopes for the Static case. The effect is smaller than predicted but is in the predicted direction. Recall, however, that Backer's hypothesis predicts that the slopes for the Dynamic case should rise quite dramatically. In fact, as the duration gets longer, the slopes for the Static and Dynamic conditions appear to converge. There is no evidence that Dynamic slopes rise to twice the Static slopes even when the stimulus is presented for 3 seconds.

4.6. Implications of Memory-free Search

The title of this paper refers to "two surprises". The possibility of memory-less search is the first of these surprises. Before turning to the second, it is worth considering some of the implications of memory-less search for our understanding of the deployment of attention.

1) At the most basic level, memory-less changes our view of the deployment of attention. We had thought it was relatively orderly. Perhaps order is expensive and perhaps reality is more anarchic, based on a simple, rapid strategy that avoids the overhead of tagging checked locations.

2) If rejected distractors are entirely unmarked, models like Guided Search would develop a problem with perseveration. Attentional deployment is biased toward the fovea (5; 64). The standard account allows attention to work its way toward a peripheral target by rejecting and marking more central distractors and then moving outward. If there is no such marking, why doesn't attention get stuck at the fovea or on the brightest or the most salient stimulus? One possibility is that there is some limited memory, perhaps a memory for the positions of the last one or two distractors. It is unclear that limited memory of this sort would have been detected in the experiments reported here. Incomplete memory has been suggested in other search contexts (e.g. 1).

3) The rate of attentional deployment in the standard models is estimated by doubling the target present slope. Thus, the standard slopes of 20-30 msec for inefficient search, implies a rate of one item every 40-60 msec. If search is memory-free, the rate is estimated directly from the target present slope, making it twice as rapid. There are investigators who have theoretical and empirical difficulty with serial selection at a rate of 40-60 msec/item because they think that attentional deployment requires several much slower steps (e.g. 12; 53). A rate of 20-30 msec/item would be even more challenging.

4) Parallel models of attention would also be disturbed by this memory-free finding. In a standard parallel model, information accumulates at each location about the likelihood of target presence. The Dynamic condition renders this accumulation function, if it were available, irrelevant. How then is it possible to search with the same efficiency in Dynamic and Static cases? These results would seem to require a parallel model that analyzes multiple, independent snapshots of the search display.

5. POST-ATTENTIVE VISION: THE SECOND SURPRISE

5.1. The Roles of Selective Attention in Object Recognition.

Earlier in this paper, it was asserted that deployment of attention to an object is a prerequisite to the recognition of that object. Why should that be the case? Selective attention serves two roles in the perception of objects. First, attention is required for the proper binding of features in objects. Prior to the arrival of attention, features of an object are not well bound to each other (47). As an illustration, see Figure Eleven.

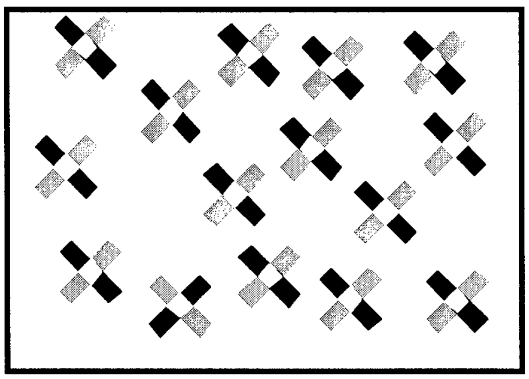


Figure Eleven: Find the black line, tilted to the right.

This is a conjunction search, logically similar to the color X orientation search shown in Figure 4. In that case, guidance from preattentive color and orientation information could lead to efficient search. Here, however, guidance fails because each "X" is treated as an object with the features "black" and "gray" and "left" and "right". Prior to the arrival of attention, the relationship of features to each other within an object is unclear. The features are "bundled" with the object but they are not "bound" (59).

In its second role in object recognition, attention controls traffic through a tight bottleneck between the visual representation of an object and its representation in memory. Recognition of a visual object requires three things. First, there must be a visual object to see and recognize. Second, there must be a representation of that object in memory. Otherwise, the observer cannot know the identity of the object. the observer would be agnosic. Finally, there must be a link between the visual and memorial representations. This notion of a link is critical. An observer might be seeing a cow and thinking of a car. We would not want this observer to 'recognize' the cow as a car. Hence, it is not enough for the two representations to coexist in time. They must be linked.

5.2. Post-attentive vision and Repeated Search

We have found that the number of links that can be maintained at any one time is very small - perhaps as small as one. The prime evidence for this conclusion comes from experiments using a "Repeated Search" paradigm in which observers search multiple times through the same set of stimuli. This is illustrated in Figure Twelve. The capital letters remain present throughout a series of N repeated searches. They do not flicker. They are not masked in anyway. Only the letter at the center changes, indicating the target for the current search. Thus, in Figure Twelve, the observer searches first for the letter 'f', next for a 'b', and so on.

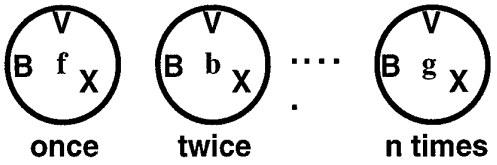


Figure Twelve: The Repeated Search paradigm. Observers, search over and over through the same, unchanging display. In this case, the display is the letters "B", "V", and "X".

We know from prior experience that the first search through these letters will be inefficient. It appears that observers must search from item to item until they find the target or, in the example shown here, until they are convinced that an "f" is not present. Search is inefficient because each letter is recognized only when attention is directed to it.

The critical question in Repeated Search concerns the fate of the effects of attention on an object after attention has been directed elsewhere. If attention allows the binding of features and the linking of visual to memorial representations, does that binding and linking survive when attention departs? The Repeated Search paradigm provides a way to answer this question. If binding and linking survive, then multiple links will be built connecting vision and memory. Eventually, all items in the display will be recognized at the same time. If the observer is then asked about an element in the display, that request will activate the node in memory. That node in memory will be linked to the visual stimulus and the observer should be able to respond, "yes", without a search. That is, RT should no longer depend on set size because the other items in the display should be irrelevant. If, on the other hand, links do not accumulate, then an inefficient search will be required each time a new target probe is presented.

5.2.1. Methods

We have performed repeated search experiments with a wide range of stimuli including letters (as shown in Fig. 12), novel objects, and 'real' objects. Details can be found in Wolfe et al. (63). Here, we will illustrate the basic result with an experiment that used conjunction stimuli of the sort shown in Figure 13.

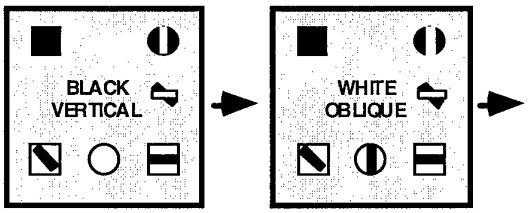


Figure Thirteen: A Repeated Search task. Observers look for the target defined by the words at the center of the display. The surrounding search array does not change.

The actual stimuli were conjunctions of color and form/orientation. Conjunction search of this sort, with variable targets and many types of distractors, is inefficient - at least on the first trial. In this experiment, observer's searched through the same display five times. One hundred sets of five trials were run at each of two set sizes, allowing us to compute slopes of the RT x set size function for each repetition.

In addition to the Repeated Search condition, an Unrepeated Search condition was run. In this case, the items changed on each trial. This condition provides a baseline for comparison. No links can be built up over repetitions in this case because no stimuli are repeated.

5.2.2. *Results*

Figure Fourteen shows the results for this experiment. The upper panel shows mean RTs as a function of repetition. The lower panel shows the slopes of the RT x set size functions.

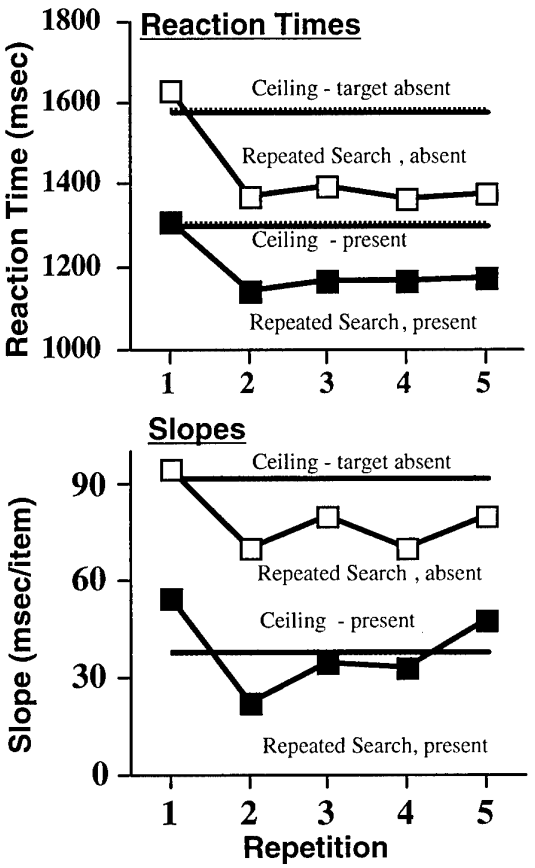


Figure Fourteen: RT and slope results of Repeated Search for conjunction targets compared to Unrepeated search for the same type of targets.

Note that, for the control "Ceiling" condition, repetition is meaningless. The stimuli are new on each trial. Accordingly, the single mean RT and slope values are plotted as straight lines, constant across the repetition variable that is of interest in the Repeated Search condition. In Repeated Search, there is some apparent improvement in the RTs from the first to the second repetition of the stimuli. However, other experiments in this series show that to be an effect of the masking of the probe words at the center by the surrounding visual stimuli (63). That masking is present on all trials in the Ceiling condition and, accordingly, the Ceiling RTs are very similar to the repetition one, Repeated Search RTs.

Turning to the slopes, we again see a hint of an improvement, mostly on the target absent trials. However, there are two important points to be made. First, even after any improvement, the search remains very inefficient. There is no hint that repeated search through these stimuli has produced the efficient search predicted if multiple items are simultaneously recognized - simultaneously linking their visual and memorial representations. Second, the target present slopes are essentially the same in the Repeated Search and Ceiling conditions, indicating that repeated search through the stimulus did not lead to the development of any representation that could facilitate search.

5.2.3. *Discussion*

We have repeated this basic finding with letters and objects; always obtaining the same general pattern of results. If search is inefficient on first exposure to a stimulus, it remains inefficient after repeated searches through that stimulus. In many cases, there is no significant change in the slope of RT x set size functions or in error rates. (Wolfe, et al., 1999). Concerned that five repetitions might be too few, we had subjects search 350(!) times through the same sets of three or five letters. Even in this extreme case, search efficiency did not improve in the Repeated Search condition.

6. GENERAL DISCUSSION

6.1. The Role of Memory in Visual Search

Two findings have been highlighted in this paper. First, the results from the Dynamic Search experiments indicate that rejected distractors are not marked during the course of a visual search. Second, the work with Repeated Search shows that search for ever changing targets does not become more efficient with repeated search through the same display. This can sound like some sort of 'attentional stupidity' or like a denial of any role for memory in visual search. Such a position would be not only counter-intuitive but wrong. Starting with the dynamics of a single search, while subjects may not keep track of rejected distractors, they must keep track of accepted targets. That, after all, is the purpose of the search. Brad Gibson and his colleagues (personal communication) have illustrated this point in a simple extension of our work. They had subjects discriminate between displays containing one or two targets. The displays could be either static or dynamic. The static case was easy. The dynamic case was virtually impossible. In the static case, subjects could find and retain the first target and then proceed to search for the second. In the dynamic case, this was impossible (given that the targets were identical. With two different targets, the results would be different.) Our claim that "visual search has no memory" is a claim of amnesia for the course of the search, not for its consequences.

The Repeated Search, post-attentive experiments are open to similar misinterpretation. It would be foolish to deny that subjects learn and remember something about the displays in repeated search tasks. After multiple searches through one display, the contents of that display are committed, at least, to some short term memory. Indeed, we compared performance on the Repeated Search tasks to performance on memory search tasks. For example, in the experiment where subjects searched through the same letters 350 times, we also included a memory search condition in which they committed letters to memory and then searched that memory 350 times. Efficiency (slope) and RT were actually somewhat faster in the absence of the visual stimulus though errors were somewhat higher.

The conclusion is that the presence of the visual stimulus conveys no benefit in Repeated Search performance.

As in the Dynamic Search experiments, this does *not* mean that subjects do not learn the locations of targets. Once you learn that bathroom is around the corner to the left, you do not have to search randomly for it. In a search paradigm, Chun and Jiang (7) have shown that subjects can learn the layout of meaningless search displays if they are repeated. That learning seems to be implicit. That is, they behave as if they remember the displays, even in the absence of any ability to explicitly recognize them among novel displays.

Our results do not deny an ability to remember displays. They merely show that the physical presence of the display does not allow a short-cut around the limited-capacity of search through that memory.

6.2. Why does visual search have no memory? Implications for artificial search mechanisms.

If one were building a search device from the ground up, one might think that it should be constructed with characteristics other than those described here for human visual search. Why not build a search mechanism that marked rejected distractors and, thus, gained efficiency over a mechanism that did not? Why not build a visual system that could have multiple links between visual and memorial representations of objects? Of course, answers to such questions are speculative. However, when Nature picks an apparently inferior way to perform a task, we may guess that the superior method was too expensive.

In the case of the marking of rejected distractors, we know that there are mechanisms of inhibition that serve to keep attention away from previously attended items(20). The most prominent of these is "inhibition of return" (IOR - 33; 39). Another apparently different mechanism has been dubbed "visual marking" (41; 54). Why should visual search use these mechanisms to avoid resampling of rejected distractors? In this case, the cost may be time. Distractors in visual search are being rejected at a rate of about 30-50 Hz (20 - 30 msec/item). These inhibitory mechanisms seem to require an order of magnitude more time (e.g. 22). By the time that this sort of inhibition could be applied, search might well be over.

Interestingly, the time course of inhibition is similar to the time of saccadic eye movements (3-4Hz). Klein and MacInnes (19) have new evidence that IOR might aid search, not by marking rejected distractors but by preventing the eyes from returning to previously fixated locations. One can imagine covert deployments of attention working cooperatively with slower, overt movements of the eyes. The eyes go to a location. Attention randomly samples 6 - 10 objects, probably in the neighborhood of fixation. This sampling is done without marking distractors but when the eyes move again, IOR prevents the same location from being the target of another eye movement. In longer searches, this could act to limit the amount of resampling of rejected distractors.

The cost of multiple links between vision and memory seems qualitatively different. It may be very hard to prevent 'cross talk' if multiple links are present. If the scene contains a car and cow and memory contains representations of a car and a cow, it is important not to attempt to drive the cow or milk the car. Selective attention may be the price we pay for accurate recognition. Kevin O'Regan (30) suggests that we can afford to pay this cost because the world serves as its own memory. Ignoring the odd case of laboratory displays with randomly changing items, the world is a fairly stable place. A cow and a car, if present at one instant, are likely to be present at the next. Even if they move, they move on trajectories that are predictable in the short term. Thus, rather than simultaneously recognizing multiple objects, we can maintain a single link from vision to memory, secure in the knowledge that we can use visual search to quickly reacquire an object if we need it. At 30-50 objects/sec we can afford to do a lot of selection.

7. CONCLUSION

There may be other ways to build a search mechanism. Perhaps slow deployment of attention would work if combined with an ability to simultaneously recognize multiple objects. However, humans and, we presume, other animals have done well with a fast but sloppy selection mechanism and a narrow channel between vision and memory.

8. REFERENCES

1 Arani, T., Karwan, M. H., & Drury, C. G. (1984). A variable-memory model of visual search. *Human Factors, 26*(6), 631-639.

2 Bauer, B., Jolicœur, P., & Cowan, W. B. (1996). Visual search for colour targets that are or are not linearly-separable from distractors. *Vision Research, 36*(10), 1439-1466.

3 Bergen, J. R., & Julesz, B. (1983). Rapid discrimination of visual patterns. *IEEE Trans on Systems, Man, and Cybernetics, SMC-13*, 857-863.

4 Bilsky, A. A., & Wolfe, J. M. (1995). Part-whole information is useful in size X size but not in orientation X orientation conjunction searches. *Perception and Psychophysics, 57*(6), 749-760.

5 Carrasco, M., & Frieder, K. S. (1997). Cortical mgnification neutralizes the eccentricity effect in visual search. *Vision Research, 37*(1), 63-82.

6 Cave, K. R., & Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology, 22*, 225-271.

7 Chun, M., & Jiang, Y. (1998). Contextual cuing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*, 28-71.

8 Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in RSVP. *Journal of Experimental Psychology : Human Perception & Performance, 21*(1), 109-127.

9 Cohen, A., & Ivry, R. B. (1991). Density effects in conjunction search: Evidence for coarse location mechanism of feature integration. *J. Exp. Psychol. Human Perception and Performance, 17*(4), 891-901.

10 D'Zmura, M. (1991). Color in visual search. *Vision Research, 31*(6), 951-966.

11 Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 433-458.

12 Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attention dwell time in human vision. *Nature, 369*(26 May), 313-314.

13 Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjunctively defined targets. *J. Exp. Psychol: Human Perception and Performance, 10*, 32-39.

14 Enns, J. T., & Rensink, R. A. (1990). Scene based properties influence visual search. *Science, 247*, 721-723.

15 Grossberg, S., Mingolla, E., & Ross, W. D. (1994). A neural theory of attentive visual search: Interactions of boundary, surface, spatialand object representations. *Psychological Review, 101*(3), 470-489.
16 Harris, J. R., Shaw, M. L., & Bates, M. (1979). Visual search in multicharacter arrays with and without gaps. *Perception and Psychophysics, 26*(1), 69-84.
17 Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(Aug 6), 575-577.
18 Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature, 334*, 430-431.
19 Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science, in press*(July).
20 Klein, R. M., & Taylor, T. L. (1994). Categories of cognitive inhibition with reference to attention. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language.* . New York: Academic Press.
21 Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897-1916.
22 Mackeben, M., & Nakayama, K. (1987). Sustained and transient aspects of extra- foveal visual attention. *Invest. Ophthalmol. Vis. Sci. (suppl), 28*, 361.
23 McLeod, P., Driver, J., & Crisp, J. (1988). Visual search for conjunctions of movement and form is parallel. *Nature, 332*, 154-155.
24 McLeod, P., Driver, J., Dienes, Z., & Crisp, J. (1991). Filtering by movement in visual search. *J. Experimental Psychology - Human Perception and Performance, 17*(1), 55-64.
25 Moore, C. M., & Wolfe, J. M. (2000). Getting beyond the serial/parallel debate in visual search: A hybrid approach. In K. Shapiro (Ed.), *The Limits of Attention: Temporal Constraints on Human Information Processing* . Oxford: Oxford U. Press.
26 Moraglia, G. (1989). Display organization and the detection of horizontal lines segments. *Perception and Psychophysics, 45*, 265-272.
27 Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320*, 264-265.
28 Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320*, 264-265.
29 Nothdurft, H. C. (1993). Faces and facial expression do not pop-out. *Perception, 22*, 1287-1298.
30 O'Regan, K. (1992). Solving the 'real' mysteries of visual perception. The world as an outside memory. *Canadian J. of Psychology, 46*, 461-488.
31 Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception and Psychophysics, 41*, 191-201.
32 Posner, M. I. (1980). Orienting of attention. *Quart. J. Exp. Psychol., 32*, 3-25.
33 Posner, M. I., & Cohen, Y. (1984). Components of attention. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X* (pp. 55-66). Hillside, NJ: Erlbaum.
34 Quinlan, P. T., & Humphreys, G. W. (1987). Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception and Psychophysics, 41*, 455- 472.
35 Ratcliff, R. (1978). A theory of memory retrieval. *Psych. Preview, 85*(2), 59-108.
36 Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *J. Experimental Psychology: Human Perception and Performance, 18*(3), 849-860.
37 Shapiro, K. L. (1994). The attentional blink: The brain's eyeblink. *Current Directions in Psychological Science, 3*(3), 86-89.
38 Suzuki, S., & Cavanagh, P. (1995). Facial organization blocks access to low-level features: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance, 21*(4), 901-913.
39 Taylor, T. L., & Klein, R. M. (1999). On the causes and effects of inhibition of return. *Psychonomics Bulletin and Review, 5*(4), 625-643.
40 Theeuwes, J., & Kooi, J. L. (1994). Parallel search for a conjunction of shape and contrast polarity. *Vision Research, 34*(22), 3013-3016.
41 Theeuwes, J., Kramer, A. F., & Atchley, P. (1998). Visual marking of old objects. *Psychonomic Bulletin and Review, 5*(1), 130-134.
42 Tipper, S. P., Weaver, B., & Watson, F. L. (1996). Inhibition of return to successively cued spatial locations: Commentary on Pratt and Abrams (1995). *Journal of Experimental Psychology: Human Perception & Performance, 22*(5), 1289-1293.
43 Townsend, J. T. (1971). A note on the identification of parallel and serial processes. *Perception and Psychophysics, 10*, 161-163.
44 Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. *J. Mathematical Psychology, 14*, 219-239.
45 Townsend, J. T. (1990). Serial and parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1*, 46-54.
46 Treisman, A. (1993). The perception of features and objects. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control.* (pp. 5-35). Oxford: Clarendon Press.
47 Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6*, 171-178.
48 Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.
49 Treisman, A., & Sato, S. (1990). Conjunction search revisited. *J. Exp. Psychol: Human Perception and Performance, 16*(3), 459-478.
50 Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of seperable features. *J. Exp. Psychol. - General, 114*, 285-310.
51 Tsotsos, J. K., Culhane, S. N., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*, 507-545.
52 von der Heydt, R., & Dursteler, M. R. (1993). Visual search: Monkeys detect conjunctions as fast as features. *Investigative Ophthalmology and Visual Science, 34*(4), 1288.
53 Ward, R., Duncan, J., & Shapiro, K. (1996). The slow time-course of visual attention. *Cognitive Psychology, 30*(1), 79-109.
54 Watson, D. G., & Humphreys, G. W. (1997). Visual marking: Prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological Review, 104*(1), 90-122.

55 Wolfe, J. M. (1992). "Effortless" texture segmentation and "parallel" visual search are *not* the same thing. *Vision Research, 32*(4), 757-763.
56 Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1*(2), 202-238.
57 Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13-74). Hove, East Sussex, UK: Psychology Press Ltd.
58 Wolfe, J. M. (1998). What do 1,000,000 trials tell us about visual search? *Psychological Science, 9*(1), 33-39.
59 Wolfe, J. M., & Bennett, S. C. (1997). Preattentive Object Files: Shapeless bundles of basic features. *Vision Research, 37*(1), 25-44.
60 Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the Feature Integration model for visual search. *J. Exp. Psychol. - Human Perception and Perf., 15*, 419-433.
61 Wolfe, J. M., & Franzel, S. L. (1988). Binocularity and visual search. *Perception and Psychophysics, 44*, 81-93.
62 Wolfe, J. M., & Gancarz, G. (1996). Guided Search 3.0: A model of visual search catches up with Jay Enoch 40 years later. In V. Lakshminarayanan (Ed.), *Basic and Clinical Applications of Vision Science* (pp. 189-192). Dordrecht, Netherlands: Kluwer Academic.
63 Wolfe, J. M., Klempen, N., & Dahlen, K. (1999). Post-attentive vision. *Experimental Psychology: Human Perception and Performance, in press.*
64 Wolfe, J. M., O'Neill, P. E., & Bennett, S. C. (1998). Why are there eccentricity effects in visual search? *Perception and Psychophysics, 60*(1), 140-156.
65 Wolfe, J. M., & Pokorny, C. W. (1990). Inhibitory tagging in visual search: A failure to replicate. *Perception and Psychophysics, 48*, 357-362.
66 Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel / serial visual search. *J. Experimental Psychology: Human Perception and Performance, 23*(1), 244-262.
67 Zohary, E., & Hochstein, S. (1989). How serial is serial processing in vision? *Perception, 18*, 191-200.

COMPUTATIONAL MODELS FOR SEARCH AND DISCRIMINATION: AN INTEGRATED APPROACH

Anthony C. Copeland and Mohan M. Trivedi
Computer Vision & Robotics Research (CVRR) Laboratory
Electrical and Computer Engineering Department
University of California, San Diego
La Jolla, CA 92093-0407
Phone: (619) 822-0002
E-mail: trivedi@swiftlet.ucsd.edu
WWW: http://swiftlet.ucsd.edu/

1. SUMMARY

This paper presents an experimental framework for evaluating metrics for the search and discrimination of a natural texture pattern from its background. Such metrics could help identify preattentive cues and underlying models of search and discrimination, and to evaluate and design camouflage patterns and automatic target recognition systems. Human observers were asked to view image stimuli consisting of various target patterns embedded within various background patterns. These psychophysical experiments provided a quantitative basis for comparison of human judgments to the computed values of target distinctness metrics. Two different experimental methodologies were utilized. The first methodology consisted of paired comparisons of a set of stimuli containing targets in a fixed location known to the observers. The observers were asked to judge the relative target distinctness for each pair of stimuli. The second methodology involved stimuli in which the targets were placed in random locations unknown to the observer. The observers were asked to search each image scene and identify suspected target locations. Using a prototype eye tracking testbed, the Integrated Testbed for Eye Movement Studies, the observers' fixation points during the experiment were recorded and analyzed. For both experiments, the level of correlation with the psychophysical data was used as the basis for evaluating target distinctness metrics. Overall, of the set of target distinctness metrics considered, a metric based on a model of image texture was the most strongly correlated with the psychophysical data.

Keywords: target detection, human visual search, discrimination, eye tracking, target signature metrics, image texture

2. Introduction

This paper deals with the issue of the development and assessment of useful computational models and quantitative metrics for integrated search and discrimination tasks. The approach is experimental in nature, where psychophysical data provides the guidance and support for comparative assessment of various metrics. This research is performed in the overall context of search and detection of camouflaged targets in natural scenes. **Figure 1** illustrates an example of such a scene, where a human observer or a machine vision system may be required to look for and detect military targets, such as a tank. This scenario is quite general. The associated problems provide a number of interesting research issues in computational vision. For example, what is the underlying model for integrated search and discrimination? What preattentive cues affect search or discrimination? How can we evaluate the relative ease or difficulty of an observer attempting to locate a selected camouflage pattern in a natural scene? How can we design the most effective camouflage pattern for a naturally textured scene? How can we rank the capabilities of automatic target recognition systems in relative terms?

Figure 1: An illustration of camouflaged targets in a natural scene.

In this paper we describe our efforts directed toward the resolution of these kinds of questions. We restrict our investigation to only textured patterns and static images. Issues related to color, range (or depth), and motion, important as they definitely are, cannot be examined in the limits of the scope of our research. We do believe that the overall experimental framework will be of utility and value for studies involving other cues.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

The ultimate goal of this line of research is the development of a robust and quantitative means for characterizing the signature strength of a target in a sensed image. The signature strength measurement should be closely correlated to the ease or difficulty of a human observer attempting to detect it [1]. In this context, the signature strength of a target is equivalent to the distinctness of the image pattern representing the target from the pattern of its specific background. Metrics that are successful at measuring perceived target distinctness would be a key component of a computational model of human visual target acquisition [2]. Such a model could form the basis of an automatic target recognition system for autonomous robot sensing or military weapons applications [3]. It could also serve to improve the assessment of military camouflage patterns and the development of more effective ones [4].

For the purpose of defining the scope of this research, we will consider human target acquisition to involve target detection followed by target recognition. The detection task is that which establishes the existence and location of an object. Recognition is the task of determining the characteristics of the object which indicate its identity, such as its size, shape, etc. Further, we will consider target detection to consist of the combination of the individual tasks of search and discrimination. Search is the process of locating areas of a scene in which to direct our attention. Discrimination is the process of segregating a potential object from its immediate background. This approach is very similar to the conclusion of O'Kane *et al.* [5]. In this paper, we are concerned with the target detection task, comprising search and discrimination, without considering recognition.

We conducted two different types of psychophysical experiments to generate quantitative measurements of perceived target distinctness for comparison to various target distinctness metrics. The first type of experiment involved paired comparisons of image stimuli that contain a target pattern embedded in a background pattern, in a constant location known to the observers. The patterns consisted of various textures extracted from images of natural scenes. For every stimulus, the target field consisted of a square shape of a constant size. We say that this experiment is a study of pure discrimination, since there is no search or recognition involved. For each pair of stimuli, the observer was required to select which of the pair possesses a target that is more distinct. By combining the decisions from a number of observers, it was possible to estimate numerical scale values for the relative levels of perceived target distinctness in the stimuli. These psychological scale values were compared to the computed values of target distinctness metrics. The second type of experiment utilized image stimuli that contain several target patterns embedded in a background scene, in random locations unknown to the observers. In this experiment, the observer needed to perform both search and discrimination. As the observer searched the scene for targets, his eye fixation points were determined by processing video of the observer's eye. The fixation point data from several observers were used to compute various statistics for each target indicating how easily the observers located it, including the likelihood the target was fixated or identified and the time required to do so. These computed statistics served as another quantitative basis for evaluating the relative effectiveness of target distinctness metrics at representing perceived target distinctness.

2. TARGET DISTINCTNESS METRICS

In some previous experiments [6, 7, 8, 9], we have observed three major perceptual cues that humans tend to utilize in judging target distinctness. These cues can roughly be called *contrast, texture differences, and boundary strength*. There are certainly many other possible perceptual cues, but these three seem to be the strongest. In this section, we discuss some specific metrics that attempt to measure the strengths of these three perceptual cues for a particular target and its local background.

2.1. Measuring Contrast

Contrast is typically measured with first-order metrics, ones that can be computed solely from the histograms of the target and local background fields [5]. A histogram is considered a first-order probability distribution since it can be calculated by considering the gray levels of pixels individually (one at a time). Statistics calculated from a histogram are capable of characterizing the overall brightness and variance of the patterns. Probably the earliest target distinctness metric is the

$$\Delta T = \left|\mu_t - \mu_b\right|,$$

area-weighted average ΔT [10], which is simply the difference between μ_t and μ_b, the computed mean gray levels of the target and background fields:

The Doyle ΔT [5] incorporates the computed standard deviations of the target and background fields, σ_t and σ_b:

Eff_POT, an abbreviation for "effective pixels on target," is computed as the number of pixels in the target pattern which

$$Doyle = \sqrt{\left(\mu_t - \mu_b\right)^2 + \left(\sigma_t - \sigma_b\right)^2}.$$

have a gray level that differs from the mean gray level of the local background pattern by more than two standard deviations of the background histogram. This metric has shown promise, especially when combined with the Doyle [5].

2.2. Measuring Texture Differences

The texture cue has been successfully measured with second-order metrics, ones computed from the gray level cooccurrence (GLC) probability distributions of the target and the background [7, 11, 12]. After Bela Julesz made the important conjecture about the role of second-order statistics in human texture discrimination, GLC models have found many useful applications in machine vision [13]. In several studies to compare the relative power of various texture analysis techniques to perform texture discrimination, GLC matrices generally outperformed other methods [14, 15, 16]. GLC's have also been used for object detection [17], scene analysis [18], as well as texture synthesis [19, 20, 21]. Other studies have demonstrated the wealth of texture information contained within GLC's [22, 23, 24].

A GLC probability distribution is calculated by considering the gray levels of pixels in pairs (two at a time), capturing

information about the spatial relationships between pixels. As such, GLC probabilities are often used as a model of image texture. One second-order metric that has shown great promise is average cooccurrence error (ACE) [7].

It is defined as

$$ACE = \frac{1}{\tau_{NGLC}} \sum_{\Delta \in D} \sum_{i=0}^{G-1} \sum_{J=0}^{G-1} \left| P_t(i, j \mid \Delta) - P_b(i, j \mid \Delta) \right|,$$

where τ_{NGLC} is the total number of displacement vectors in the set D of vectors in the texture model, G is the number of possible gray levels, $P_t(i, j|\Delta)$ is the joint probability of a pixel of gray level i and a pixel of gray level j given the displacement vector $\Delta=[\Delta_x\ \Delta_y]$ for the target pattern, and $P_b(i, j|\Delta)$ is the corresponding joint probability for the background pattern. For computing this metric, we normally consider all possible displacements of up to a maximum of $\tau_{NX}=\tau_{NY}=8$ pixels, yielding a total of $\tau_{NGLC}=2\tau_{NX}\tau_{NY}+\tau_{NX}+\tau_{NY}=144$ displacements. If the original image is quantized to 256 gray levels, the pixel values in the target and background regions are reduced to $G=8$ possible gray levels for computation of the model. Since each of the 144 GLC matrices in the texture models is of size $G \times G$, using a full $G=256$ gray levels produces a data structure that is prohibitively large.

2.3. Measuring Boundary Strength

The third class of target distinctness metrics we considered consists of metrics, which attempts to quantify target/background boundary strength. Even if a target's texture pattern is very similar to the texture of its local background, discontinuities along the target/background boundary can still serve as a perceptual cue [25]. One way to measure this is to compute the average contrast between the pixels lying on either side of the target/background boundary. For a single point i along a boundary, the contrast is

$$c(i) = \left| p_t(i) - p_b(i) \right|,$$

where $p_t(i)$ is the gray level of the pixel just on the target side of the boundary and $p_b(i)$ is the gray level of the adjacent pixel just on the background side. For a target field that is a rectangular lattice of pixels, the lengths of the boundaries are $n_{top} = n_{bottom} = n_{horiz}$ and $n_{left} = n_{right} \equiv n_{vert}$. The average contrast for one boundary (such as the top boundary) is

$$C_{top} = \frac{1}{n_{horiz}} \sum_{i=1}^{n_{horiz}} c(i),$$

where i is just a summation index for the boundary points.

Then the average boundary strength (ABS) for the whole target is:

$$ABS = \frac{n_{horiz}\left(C_{top} + C_{bottom}\right) + n_{vert}\left(C_{left} + C_{right}\right)}{2n_{horiz} + 2n_{vert}}.$$

For a target field that is a perfect square, such as in our stimulus images, we have $n_{horiz} = n_{vert} \equiv n_{bound}$. In this case, the equation for ABS reduces to

The ABS measure does not take into account the values of any pixels that do not lie adjacent to the target/background

$$ABS = \frac{1}{4n_{bound}} n_{bound}\left(C_{top} + C_{bottom} + C_{left} + C_{right}\right)$$

$$= \frac{1}{4}\left(C_{top} + C_{bottom} + C_{left} + C_{right}\right).$$

boundary. However, a target/background boundary that has a high value for ABS may not be very distinct if it is embedded in a region that already is characterized by a large amount of contrast. To take into account the contrast of the entire region, we use relative average boundary strength (RABS):

where n_{region} is the number of adjacent (either vertically or horizontally) pixel pairs within the target field or in the

$$RABS = \frac{ABS}{\frac{1}{n_{region}} \sum_{i=1}^{n_{region}} c(i)},$$

background near the target. Essentially, RABS is the ratio of the average contrast along the target/background boundary to the average contrast between adjacent pixels in the vicinity.

3. THE ITEMS TESTBED

3.1. Overview and Utility of ITEMS

This section discusses the design and implementation of ITEMS – the Integrated Testbed for Eye Movement Studies. This prototype eye tracking testbed consists of an integrated system of hardware and software which allows an experimenter to present an observer with an image displayed on a high resolution monitor and have the observer perform a visual task. **Figure 2** shows a test subject studying a displayed image scene while ITEMS tracks his eye fixation points. Using ITEMS, not only can we determine whether a particular target was identified by an observer, but also whether the target was ever fixated by the observer (even if it was not identified as being a target), how long did it take before the target was first fixated, how long the target was studied before it was identified, what search path the observer took on the way to the target, and any number of other aspects of visual search.

The hardware components of ITEMS are a Silicon Graphics Indy computer workstation with high resolution color monitor, a Sony CCD black and white video camera fitted with a 50mm lens and 5mm lens spacer, a Datacube MaxTD image processing system containing a MaxVideo~200 pipeline processor and MVME-167 CPU system controller, and an adaptable yet sturdy apparatus to which is mounted the camera as well as a helmet for restricting observer head movements. The software components of ITEMS include an X-Windows application to handle the image scene display and observer response registration for the Indy workstation, pupil centroid tracking and registration for the Datacube MaxTD, a utility for fixation point estimation and head movement adjustments, a utility for spatial calibration and

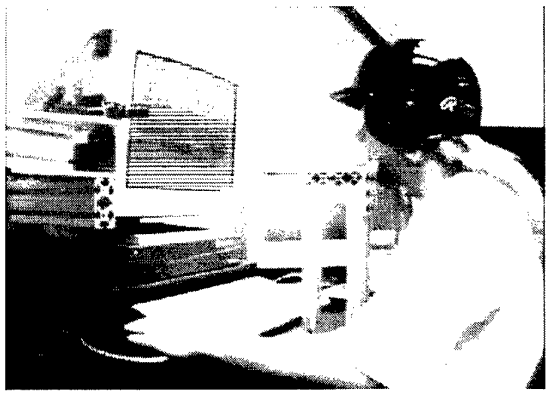


Figure 2 A test subject studying a displayed image scene while ITEMS tracks his eye fixation. points.

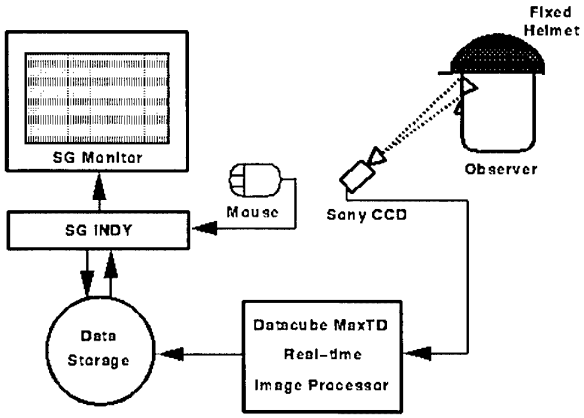


Figure 3: The interconnectedness of the

error interpolation, and another for calculation of target fixation and identification statistics.

All required images are created by the experimenter beforehand. This includes a zero point image, several calibration images, and all desired experiment image scenes. The procedure is that the zero point image is displayed first, then all calibration images in succession, the zero point image again, and then any number of sessions of experiment images, each pair of sessions separated by another presentation of the zero point image. The number of experiment sessions and number of images in each session can vary, but it has been found that five images per session with one calibration session and two experiment sessions results in a moderate ten minutes of data collection for each observer.

The zero point image is an image with one target that is located such that it is directly ahead of the observer's left eye when displayed on the monitor. The target consists of a square region of uniform gray level against a background of a different gray level. This image is used to establish a reference point to which all eye movements can be related and also to measure periodically the change in fixation point estimates that is the result of small head movements accumulating over time. This procedure is described later in Section 4.5.

Each calibration image consists of a row of square targets. The targets in all the calibration images taken together constitute an array of evenly spaced points, which are used as sample points at which to measure the error in fixation point estimates due to measurement and modeling error. As these errors will vary over different spatial locations in the display image, a number of samples are taken and then adjustments are made in fixation point estimates from an interpolation of the calibration samples from the vicinity of each estimate.

3.2. ITEMS Hardware Configuration

Figure 3 shows the interconnectedness of the various hardware components of ITEMS. Briefly, the Silicon Graphics Indy workstation is used to load each image scene stimulus from disk and display it on its high-resolution color monitor. The Sony CCD video camera sends continuous video to the Datacube MaxTD image processor, which locates, tracks, and records the pupil centroid location of the observer's left eye. The observer's head movement is restricted using a baseball-batting helmet, which is rigidly mounted to the table, using adjustable aluminum extrusion material. This material allows the helmet to be raised or lowered to accommodate different observers and also to be locked into place when the appropriate position is found. Head movement is further restricted by a chin rest.

The camera is mounted directly in front of and below the observer, just below the Silicon Graphics monitor, looking upward at the observer's left eye. This location was found to provide an adequate image of the observer's left eye and a small reference mark affixed just below the eye. This reference mark is a small, glossy black paper circle, used to distinguish eye movements from small head movements. The observer's face is illuminated with a small portable flashlight as necessary to segregate the pupil and reference mark together from the rest of the video image. The Datacube MaxTD also has a small terminal screen, which allows the experimenter to monitor the status of the image processor's eye tracking, and the video from the CCD camera is simultaneously displayed on a small monitor for the same purpose.

3.3. Image Scene Display and Observer Response Registration

Image scene display and observer target identification response registration for ITEMS is handled by the Silicon Graphics Indy workstation. The X-Windows application created for this purpose is called I_SPY. I_SPY is used to load each image scene stimulus from disk and display it on the high-resolution color monitor. In experiment mode, the observer uses mouse buttons to indicate when to display each image, when he wishes to identify a suspected target, and when he is finished searching a particular scene. In playback mode, I_SPY allows the experimenter to study the data by displaying the image scene stimuli with a cursor which moves about the images indicating the observer's fixation points over time.

3.4. Pupil Centroid Tracking and Registration

The tracking and registration of pupil centroid position is handled by the Datacube MaxTD image processing system. The procedure is to first threshold the video frame such that both the observer's pupil and the black paper circle affixed just below his eye appear as black circular blobs in the image. The resulting binary image is then subjected to a connectivity analysis, which computes the number of blobs in the image and a roundness measure for each. The roundness measure is computed by finding a best-fit ellipse for each blob, and calculating the ratio of the two axes of the ellipse. The roundness measure is used to separate the pupil and reference mark blobs from various shadow artifacts, which generally do not appear as round blobs at all. The values that are stored are the centroid differences in both x-coordinates and y-coordinates between the upper blob (the pupil) and the lower blob (the reference mark), along with the current timestamp. Thus it is only movement of the pupil relative to the reference mark that is tracked and registered. In this way, movements of the eye can be distinguished from small head movements. That is, a small head movement will result in a change of position of both the pupil and the reference mark in the camera image. Although a helmet mounted in a fixed position and a chin rest are used to restrict observer head movement, in practice there is still a bit of a small head movement even with the most cooperative observers, due to breathing, heartbeats, etc.

3.5. Eye Tracking Geometry and Fixation Point Estimation

Details of the fixation point estimation process are given in reference [26]. Briefly, the steps necessary to obtain the fixation estimate for each data sample are:

1. Extract the values for the difference in x-and y-coordinates between the pupil centroid and the reference point centroid from the data file of the pupil centroid tracking program.
2. Compare these values to the same values from the moment the observer identified the first zero point. The change is taken to be the movement of the pupil center in the camera image from the zero state.
3. From the location of the pupil center in the camera image, find its location in world coordinates using the inverse perspective transform [27], subject to the constraint that the point is known to lie on the front side of the sphere representing the eyeball.
4. Based on the location of the pupil center in world coordinates, find the intersection point of the line representing the visual axis and the plane representing the display image.
5. Find the fixation point estimate by converting the location of the intersection point from world coordinates to display image coordinates (x' and y').
6. Adjust the fixation point estimate for small head movements by subtracting the average of the error for the zero point at the beginning of the session and the one at the end of the session. For each zero point, the error is taken to be the change in fixation point estimate since the first zero point image at the beginning of the calibration session.

4. STUDYING PURE DISCRIMINATION

This section describes a psychophysical experiment designed to investigate the task of human target discrimination separate from visual search, or "pure discrimination." The image stimuli used in this experiment consisted of target patterns embedded in background patterns, in a constant location known to the observers. With such stimuli, it is unreasonable to ask observers to make absolute judgments of target distinctness because of the complex nature and wide range of criteria that could be used in such a judgment. Instead, we only asked the observers to make relative judgments of target distinctness. The image stimuli were presented in pairs, and the observers were required to select which image of each pair possesses a target that is more distinct. By combining the decisions from a number of observers, it is possible to estimate numerical scale values for the relative levels of perceived target distinctness in the stimuli. These psychological scale values were used as a quantitative basis for evaluating the relative effectiveness of our target distinctness metrics at representing perceived target distinctness. The established method for accomplishing this "psychological scaling" is the law of comparative judgment (LCJ), introduced by Thurstone [28, 29]. The LCJ is based on the postulate that if a stimulus is presented to a human subject, it excites a discriminal process, which has some value on the psychological continuum. It is also assumed that this value will not be exactly the same each time the same stimulus is presented, but rather these values will form a normal distribution along the continuum. For more information about the specific method to estimate the scale values, see reference [7].

The 15 image stimuli used in the experiment are shown in **Figure 4**. The computer environment that was developed to automate the sequential display of the image stimulus pairs and the registration and recording of subject responses is the X-based Perceptual Experiment Testbed (XPET) [6, 7]. XPET was used to present 20 observers with all 105 possible pairs of the 15 stimuli. The raw judgments were used to estimate an appropriate scale value for each stimulus. **Figure 5** shows graphically the locations of the scale values along the perceptual continuum representing target distinctness. These scale values indicate only relative amounts of target distinctness in the stimuli as judged by the observers, and have no absolute meaning. The stimulus containing the target judged least distinct was stimulus DF. This stimulus is assigned a scale value of zero, and the scale is constructed upward from that point. The stimuli containing the most distinct targets as judged by the observers were stimuli CF and CD. The sample correlation coefficient was then computed between the vector of psychological scale values and the vector of each of the computed target distinctness metrics. The results are given in Table 1. **Figure 6** shows the test images plotted with their LCJ scales and computed values for the ACE metric.

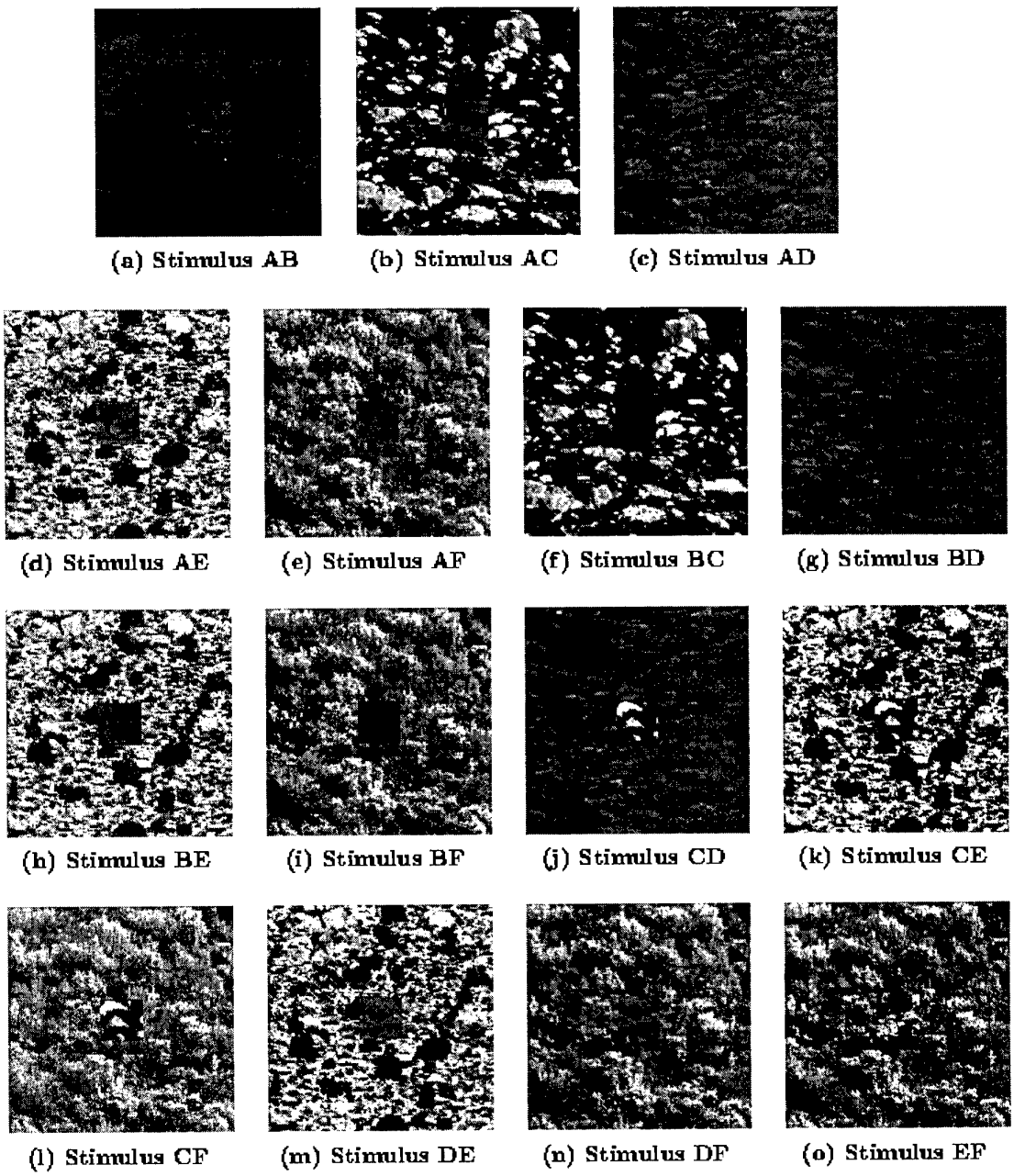


Figure 4: The 15 256 x 256 test images for the discrimination experiment.

4.1. Multivariable Linear Regression and Multiple Correlation

We now compare the psychological scale values to not one, but several variables. The single variable linear regression model is of the form

$$y = \beta_0 + x\beta_1 + \varepsilon,$$

where y is the response (dependent) variable, x is the independent variable, β_0 and β_1 are regression parameters, and ε is the error which is presumed to be normally distributed with mean of $\mu=0$ and variance of σ^2. Previously, y represented the stimulus scale value estimated from the psychophysical data and x represented any one of the image metrics that we were studying. With N stimuli in the experiment, we actually have N samples of both y and x, so the entire model is written

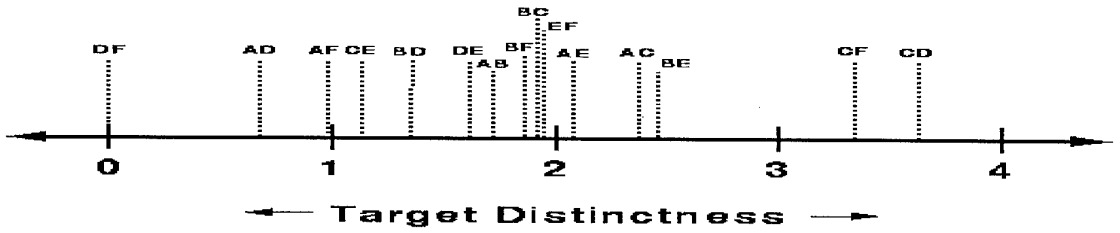


Figure 5: The relative locations of the scale values along the perceptual continuum representing target distinctness.

TABLE 1: The sample correlation coefficients (r) between the vector of stimulus scale values for perceived target distinctness and the vector of each of the target distinctness metrics.

Metric	r
Δ T	0.14
Doyle	0.66
Eff_POT	0.57
ACE	0.83
ABS	0.65
RABS	0.76

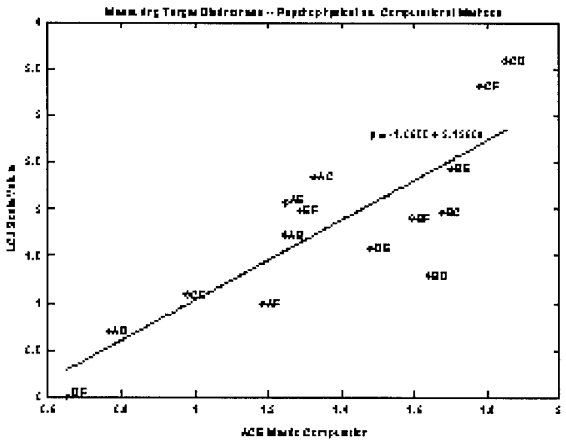


Figure 6: The test images plotted with their LCJ scales and computed values for the ACE metric.

$y = \beta_0 + x\beta_1 + \varepsilon$, where $y' = (y_1, \ldots, y_N)$ represents the N scale values $x' = (x_1, \ldots, x_N)$ represents the N computed values of the particular image metric, and $\varepsilon' = (\varepsilon_1, \ldots, \varepsilon_N)$ represents the error for each sample.

We actually have k independent variables (image metrics) interacting simultaneously. Now, the model can be written

$y = X'\beta = \varepsilon$, where the differences are that β is a $k + 1$ length vector of regression parameters and **X** is a rectangular matrix of computed image metrics with $k + 1$ rows and N columns. (Actually, the first row of **X** consists of all 1's which are dummy variables so that the additive constant parameter β_0 is included.) The least squares solution for β is given by $\hat{\beta} = (XX')^{-1}$ [30].

TABLE 2: The multiple correlation coefficients for selected pairs of metrics.

	Δ T	Doyle	Eff_POT	ACE	ABS	RABS
Δ T	-	0.72	0.59	0.83	0.65	0.78
Doyle	-	-	0.90	0.83	0.75	0.89
Eff_POT	-	-	-	0.88	0.80	0.78
ACE	-	-	-	-	0.88	0.87
ABS	-	-	-	-	-	0.80
RABS	-	-	-	-	-	-

Statistical correlation can also be extended to multiple independent variables. We previously performed a simple correlation to measure the degree of linear association between two random variables. We can now utilize multiple correlation to measure the maximum correlation between the dependent variable and a linear combination of a set of independent variables. This enables us to test the ability of various linear models for the human texture discrimination process to explain the empirical data. The multiple correlation was computed for all possible pairs of the metrics considered. This value is defined as the highest value of the correlation coefficient computed between the scale values and a linear combination of the two metrics. The results are given in Table 2.

Additionally, we can use multiple correlation to test the effectiveness of various models consisting of linear combinations of more than two metrics to predict the psychological data. For this analysis, four metrics were selected as the most promising out of the seven which were tested with pairwise correlations. These four metrics are assigned numerals 1-4 as follows:1 = Doyle, 2 = Eff_POT, 3 = ACE, and 4 = RABS. The models tested are a linear combination of all four and every possible combination of three. The results of this are given in Table3.

TABLE 3: The multiple correlation coefficients and corresponding regression parameters for selected linear combinations of metrics.

Metrics	Max Correlation	Regression Parameters				
		β_0	β_1	β_2	β_3	β_4
1,2,3,4	0.94	-1.18e+00	6.58e-02	5.48e-04	1.11e-01	3.0767e-01
1,2,3	0.91	-8.05e-01	5.34e-02	6.77e-04	8.50e-01	-
1,2,4	0.94	-1.16e+00	6.91e-02	5.57e-04	-	3.22e-01
1,3,4	0.89	-1.55e+00	4.03e-02	-	6.00e-01	4.37e-01
2,3,4	0.89	-1.20e+00	-	3.38e-04	1.51e+00	2.06e-01

In the table, the second column lists the value of the maximum correlation coefficient computed between the scale values and the linear combination of metrics in the first column. The remaining columns list the values of the regression parameters for which the model yields the maximum correlation value, corresponding to the $k + 1$ β parameters in $\hat{\beta} = (XX')^{-1}$. In each case, the value listed in the β_0 column is the value of the additive constant parameter in the linear model for the optimum case. The values of these regression parameters do not absolutely indicate the relative importance of each metric in the model, since they provide both weighting and normalizing of the metrics. They are included simply to illustrate that although the maximum correlations for the models are rather high, their eventual utility depends on the proper selection of values for several parameters.

When the metrics were considered two at a time, the highest correlation (0.90) was obtained for a linear combination of the Doyle metric with the Eff_POT metric. These two metrics were also found, in a previous experiment performed at the U.S. Army Night Vision and Electronic Sensors Directorate, to be the best predictors of the probability of finding low observable military targets in simulated infrared imagery [5].

When combinations of three or four metrics were considered, a correlation of 0.94 resulted for the combination of Doyle, Eff_POT, and RABS. The inclusion of the GLC-based ACE metric does not significantly improve this result. Thus, it seems that for the stimuli and resulting psychological scale values in this experiment, it is best to use a GLC-based error metric if a single metric is desired as a measure of target distinctness. However, if we allow the inclusion of multiple metrics in the model, it is best to discard the GLC-based metric and instead use the Doyle, Eff_POT, and RABS metrics. But before such a combination model can be used in practice, it will certainly be necessary to conduct further experimentation to either confirm the robustness of the regression parameters that were best for this experiment or to determine values that are the better for the particular imagery being used.

5. STUDYING INTEGRATED VISUAL SEARCH AND DISCRIMINATION PROCESS

This section describes a psychophysical experiment designed to investigate the task of human target discrimination when combined with visual search. The image stimuli used in this experiment also consisted of square target patterns embedded in background patterns, but in random locations unknown to the observers. As each observer performed a visual search of the scene for targets, his eye fixation point within the stimulus was measured by processing video of the observer's eye. By integrating search and discrimination, we can indirectly measure perceived target distinctness by measuring various statistics that indicate how easily the observers located it, including the likelihood the target was fixated or identified and the time required to do so. These computed statistics will also serve as a quantitative basis for evaluating the relative effectiveness of our target distinctness metrics at representing perceived target distinctness.

5.1. Creation of the Image Scene Stimuli

The images used in the visual search experiment were extracted from a set of natural scenes of various locations in southern California. All of the images were obtained using a Nikon 35mm camera and developed as 8 × 10 inch color enlargements. The enlargements were digitized at 120 pixels per inch using a Hewlett Packard digital scanner. The scenes include a wide variety of both terrain and vegetation conditions such as forests, mountains, fields, and deserts. Great care was taken to ensure that no man-made objects or animals appear in the scenes. The viewing perspective of each scene is such that the viewer is looking down from above, and the viewing distance varies from as close as 100m to as far as several kilometers.

Ten 800 × 1200 images were selected from the database as representative of the wide variety of possible terrain and vegetation conditions. The color images were converted to gray scale by averaging the red, green, and blue channels. These ten raw images were used to create ten stimulus images according to a random scheme. For each stimulus, one of the ten raw images was designated as the background image and another of the ten was chosen as the target image. A random number of either four, five, or six was chosen for the number of targets. Every target was a square region 48 pixels on each side. A random location was chosen for each target square, with the restrictions that no target pixels could lie within 96 pixels, or two target dimensions, of the boundaries of the image, and no target pixels could lie within 144 pixels, or three target dimensions, of another target's pixels. If a target location was chosen that did not meet these two restrictions, it was discarded and another random location was chosen. Once the number of targets and target locations for a particular stimulus were randomly chosen, the stimulus image was created by using the pixel values of the raw background image for all pixels except target pixels. The values for the target pixels were taken from the pixels in the raw target image at the corresponding locations. In this manner, we obtained a wide variety of naturally occurring target patterns against different, naturally occurring background patterns. There were a total of 52 targets in the

ten image stimuli. Four of the stimulus images are shown in the Appendix.

5.2. Conduct of the Experiment

Data was collected from a total of 12 different observers. Each observer was told that each of the image scenes contained between four and six targets each, and that every target is a square region of a specified size that contains a pattern which looks as if it doesn't belong in its location, in that it looks "unnatural" or "out of the ordinary." He was asked to identify each target as soon as he sees it, and to find as many of the targets in each image before proceeding to the next. The ten stimuli were presented to each observer in a different, randomly chosen order. Together with five calibration images and four zero point images, each observer was presented a total of 19 images in the experiment. This typically required about 10-15 minutes, during which time the observer was required to hold his head still. **Figure 7** shows the raw fixation point data from one observer for one of the stimulus images. The white cross hairs show the observer's fixation points during the display of that image at the discrete sample times, with consecutive sample points connected by a straight line to indicate the eye movement. The fixation points at each of the moments that the observer pressed the middle mouse button are shown as small white square blocks. These correspond to areas suspected by the observer to be targets.

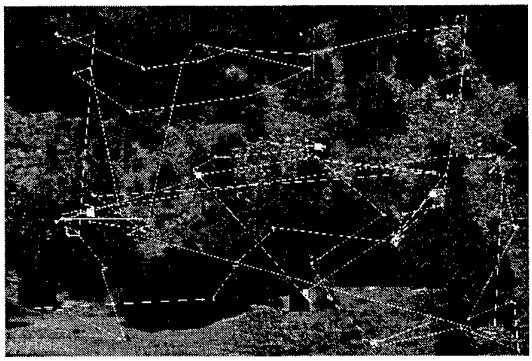


Figure 7: The raw fixation point data from one observer for one of the stimulus images. The white streaks indicate the observer's fixation points, while suspected target locations are shown as small white square blocks.

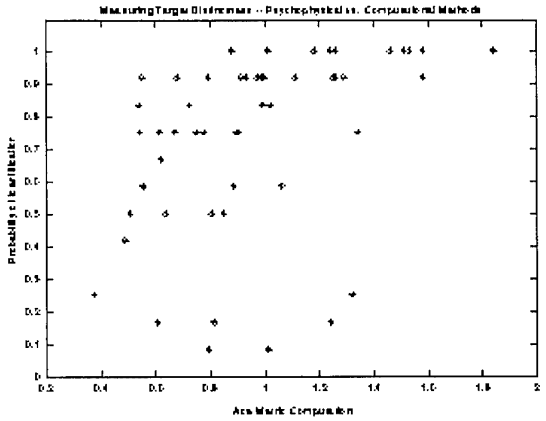
5.3. Target Fixation and Identification Statistics

The data provided by ITEMS for every observer consists of the fixation point coordinates in the display image and the corresponding timestamp for each sample, along with the timestamp and button identifier for every press of a mouse button during the session. Since the mouse button presses are the means by which the observer both controls the image display process and indicates he is fixating targets, and the locations of all targets in the image stimuli are known, this data is sufficient for computing various statistics describing the observer's search for and discrimination of the targets. When an observer is studying a particular target to decide whether it is indeed a target, his exact point of fixation will normally move about both within and just outside the target square, as he looks for cues to assist him in the decision. Thus, for the computation of these statistics, a fixation point was considered to be a fixation of a target if it was within the target square or within one and one-half target dimensions outside the target square. The statistics computed for the 52 targets in the experiment are identification probability (P_{ID}) average time to identification ($\overline{T_{<ID}}$), fixation probability (P_{fix}) average time to first fixation ($\overline{T_{<fix}}$), and average total fixation time ($\overline{T_{fix}}$). The computations of P_{ID} and P_{fix} are more properly the *likelihood* of identification and fixation for each target, as they are simply calculated as the proportion of the 12 observers that identified and fixated the target. The statistics $\overline{T_{<ID}}$ and $\overline{T_{<fix}}$ are computed as the time elapsed from the moment the image was first displayed until the observer first identified or fixated the target, averaged over only those observers that did indeed identify or fixate the target. The statistic $\overline{T_{fix}}$ is computed as the total time the observer spent fixating the target area, averaged over all 12 observers. The set of target distinctness metrics were computed for all 52 targets in the experiment. For each calculation, the background was considered to consist of all pixels not in the target square but within one target dimension. Table 4 gives the sample correlation coefficient (r) computed between the five vectors of computed target fixation and identification statistics and the vector of each of the target distinctness metrics. From Table 4, we see that for the P_{ID} and P_{fix} statistics we have $r > 0$ for all of the target distinctness metrics considered. A target that is more distinct is more likely to be fixated and/or identified. We also see that for the $\overline{T_{<ID}}$ and $\overline{T_{<fix}}$ statistics we have $r < 0$ for all of the metrics. A target that is more distinct will likely be fixated and/or identified in less time. The second-order ACE metric exhibited the strongest correlations for $\overline{T_{<ID}}$, $\overline{T_{<fix}}$, and $\overline{T_{fix}}$. For P_{ID}, ACE was just behind RABS for the most strongly correlated.

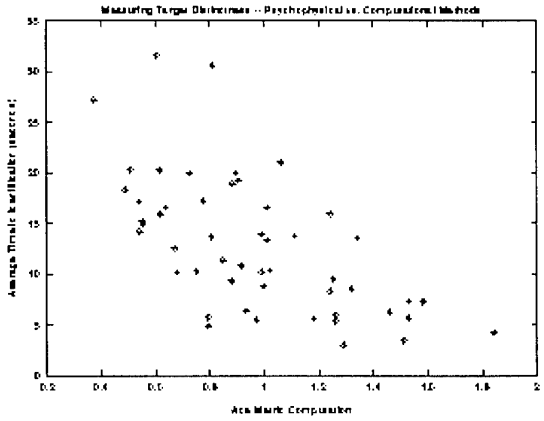
Figure 8 shows plots of the 52 targets in the search experiment, with the horizontal axis representing the computed value of the ACE metric and the vertical axis representing the P_{ID} and $\overline{T_{<ID}}$ statistics.

5.4. Analysis of the Results

For this experiment, we have found that the magnitudes of the correlations between the individual target distinctness metrics and the probability of identification (P_{ID}) were as high as 0.43 and for average time to identification ($\overline{T_{<ID}}$) were as high as 0.62. Although these values do indicate strong relationships, we must realize that there are many more variables contributing to whether an observer identifies a target and the time required to locate a target than just the



(a) P_{ID} and ACE.



(b) $\overline{T_{<ID}}$ and ACE.

Figure 8: The 52 targets in the search experiment plotted with their identification statistics and the computed values of the ACE metric.

TABLE 4: The sample correlation coefficient (r) computed between the five vectors of target fixation and identification statistics and the vector of each of the target distinctness metrics.

	P_{ID}	$\overline{T_{<ID}}$	P_{fix}	$\overline{T_{<fix}}$	$\overline{T_{fix}}$
Δ T	0.30	-0.55	0.28	-0.47	-0.32
Doyle	0.30	-0.54	0.34	-0.49	-0.29
Eff_POT	0.35	-0.43	0.24	-0.38	-0.25
ACE	0.42	-0.62	0.31	-0.56	-0.33
ABS	0.23	-0.25	0.19	-0.31	0.09
RABS	0.43	-0.43	0.35	-0.42	-0.05

distinctness of the target. It is also important to realize that even if there is a direct relationship between two variables, the computed value of a correlation coefficient between them may not be high if the relationship is not linear.

Overall, of the set of target distinctness metrics considered, the second-order GLC-based ACE metric was the most strongly correlated with the psychophysical data. Although the observers were not instructed as to what cues they were to use in making their judgments, we can surmise that the observers probably utilized some combination of differences in brightness (contrast), differences in texture, and abrupt discontinuities along target/background boundaries. Certainly differences in target and background first-order pixel probabilities are important, since they represent pattern contrast and variation. But second-order probabilities are important too, since they better represent the general concept of texture by taking into account the spatial relationships between pixels. A GLC model may be able to capture at least some of all of these variables. Second-order probabilities inherently contain first-order probabilities, in that a pattern's histogram can be obtained by summing over all rows or over all columns of one of its GLC matrices. Also, if two patterns have GLC models that are significantly different, it is apparent that a distinctly abrupt boundary is more likely if the two patterns are placed adjacent to each other.

6. CONCLUDING REMARKS

In our future studies, we wish to determine which cue is most important for each target and use a metric appropriate for that target, instead of trying to use the same metric for every target. Or, perhaps a proper weighting of the relative importance of the three perceptual cues could be determined for every target, and used to form a composite metric. Additionally, the variable of target size must be factored into the metrics. In our experiments, we also did not vary the size of the field of view, which most certainly has an effect on search times. We feel also that the spatial location of the target in the image (such as center or periphery) and global variables such as scene clutter have an effect. The model should also account for the effects of competing targets and other points of interest, as well as false alarms [31, 32].

As for the experimental methodology presented in this paper, both the pure discrimination and the search experiment allowed us to study perceived target distinctness. But the search experiment provided us with data that can be used to develop or test models describing various aspects of the search and discrimination processes, rather than only the final result. And not only do we have fixation data that include two-dimensional image coordinates, but also a third dimension of time, which will allow us to include this dimension in the model.

Besides target search and discrimination, it is apparent any study of human visual perception can benefit from measuring

the eye fixations of observers. Although we can always have an observer report his judgments of a visual stimulus, knowledge of the eye fixation points provides us with invaluable insights into the process through which the observer reached his decisions. We plan to expand the scope of our studies to include other applications which depend on human visual perception, such as advanced human-computer interfaces, adaptive videoconferencing systems, and assessment of digital display quality and television advertising effectiveness.

7. REFERENCES

1. M. M. Trivedi and M. V. Shirvaikar. Quantitative characterization of image clutter: Problems, progress, and promises. Characterization, Propagation, and Simulation of Sources and Backgrounds III}, Orlando, FL, April 1993. SPIE.
2. J. D'Agostino, W. Lawson, and D. Wilson. Concepts for search and detection model improvements. In G. C. Holst, editor, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing VIII, volume 3063, Orlando, FL, April 1997 SPIE.
3. R. Hecht-Nielsen and Yi-Tong Zhou. Vartac: A foveal active vision {ATR} system. Neural Networks}, 8(7/8):1309--1321, 1995.
4. G. W. Walker and J. R. McManamey. Characterization of natural background clutter for design of camouflage. In D. Clement and W. R. Watkins, editors, Characterization, Propagation, and Simulation of Sources and Backgrounds II}, volume 168 pages 254--264, Orlando, FL, April 1992. SPIE.
5. B. L. O'Kane, C. P. Walters, and J. D'Agostino. Report on perception experiments in support of low observables thermal performance models. Technical report, U.S. Army, Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, Feb. 1993.
6. A. C. Copeland, M. M. Trivedi, and J. R. McManamey. Evaluation of image metrics for target discrimination using psychophysical experiments. Optical Engineering}, 35(6):1714--1722, June 1996.
7. A. C. Copeland and M. M. Trivedi. Texture perception in humans and computers: Models and psychophysical experiments.In W. R. Watkins and D. Clement, editors, Targets and Backgrounds: Characterization and Representation II, volume 2742, pages 436--446, Orlando, FL, April 1996. SPIE.
8. A. C. Copeland and M. M. Trivedi. Integrated framework for developing search and discrimination metrics. In W. R. Watkins and D. Clement, editors, Targets and Backgrounds: Characterization and Representation III, volume 3062, Orlando, FL, April 1997. SPIE.
9. A. C. Copeland and M. M. Trivedi. Models and metrics for signature strength evaluation of camouflaged targets.In E. G. Zelnio, editor, Algorithms for Synthetic Aperture Radar Imagery IV, volume 3070, Orlando, FL, April 1997. SPIE.
10. J. A. Ratches. Static performance model for thermal imaging systems. Optical Engineering}, 15(6):525--530, 1976.
11. M. V. Shirvaikar and M. M. Trivedi. Developing texture-based image clutter measures for object detection. Optical Engineering}, 31:2628--2639, Dec. 1992.
12. S.R. Rotman, A. Cohen, D. Shamay, D. Hsu, and M. L. Kowalczyk Textural metrics for clutter affecting human target acquisition. In G. C. Holst, editor, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing VII, volume 2743, pages 99--112, Orlando, FL, April 1996. SPIE.
13. B. Julesz. Visual pattern discrimination. IRE Trans. Information Theory, 8(2):84--92, Feb. 1962.
14. J. S. Weszka, C. R. Dyer, and A. Rosenfeld. A comparative study of texture measures for terrain classification. IEEE Transactions on Systems, Man, and Cybernetics, 6:269--285, Apr. 1976.
15. R. W. Conners and C. A. Harlow. A theoretical comparison of texture algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2(3):204--222, May 1980.
16. P. P. Ohanian and R. C. Dubes. Performance evaluation for four classes of textural features. Pattern Recognition}, 25(8):819--833, 1992.
17. M. M. Trivedi, C. A. Harlow, R. W. Conners, and S. Goh. Object detection based on gray level cooccurrence. Computer Vision, Graphics, and Image Processing, 28:199--219, Nov. 1984.
18. C. A. Harlow, M. M. Trivedi, R. W. Conners, and D. Phillips. Scene analysis of high resolution aerial scenes. Optical Engineering}, 25(3):347--355, March 1986.
19. A. Gagalowicz and Song De Ma. Sequential synthesis of natural textures. Computer Vision, Graphics, and Image Processing, 30:289--315, 1985.
20. G. Lohmann. Co-occurrence-based analysis and synthesis of textures. In 12th IAPR International Conference on Pattern Recognition (ICPR)}, volume 1, pages 449--453, Jerusalem, Israel, Oct. 1994.
21. G. Ravichandran, E. J. King, and M. M. Trivedi. Texture synthesis: A multiresolution approach. In Proc. of the Ground Target Modeling and Validation Conf., Houghton, MI, 1994. Keweenaw Research Center.
22. A. R. Figueiras-Vidal, J. M. Paez-Borrallo, and R. Garcia-Gomez. On using cooccurrence matrices to detect periodicities. IEEE Transactions on Acoustics, Speech, and Signal Processing, 35(1):114--116, Jan. 1987.
23. J. Parkinnen, K. Selkainaho, and E. Oja. Detecting texture periodicity from the cooccurrence matrix. Pattern Recognition Letters, 11:43--50, Jan. 1990.
24. C. C. Gotlieb and H. E. Kreyszig. Texture descriptors based on co-occurrence matrices. Computer Vision, Graphics, and Image Processing, 51:70--86, 1990.

25. M. J. Muller. Texture boundaries: Important cues for human texture discrimination. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)}, pages 464--468. Miami Beach, FL, June 1986.

26. A. C. Copeland. Image Metrics for Human Search and Discrimination of Textured Targets and Backgrounds}. PhD thesis, Univ. of Tennessee, Knoxville, 1996.

27. R. C. Gonzalez and R. E. Woods. Digital Image Processing. Addison-Wesley, Reading, MA, 1992.

28. L. L. Thurstone. A law of comparative judgment. Psychological Review}, 34:273--286. 1927.

29. W. S. Torgerson. Theory and Methods of Scaling. John Wiley and Sons, New York, 1958.

30. M. S. Srivastava and E. M. Carter. An Introduction to Applied Multivariate Statistics. North-Holland, New York, 1983.

31. S. Grossman, Y. Hadar, A. Rehavi, and S. R. Rotman. Target acquisition and false alarms in clutter. Optical Engineering}, 34(8):2487--2495, Aug. 1995.

32. T. J. Doll and D. E. Schmieder. Observer false alarm effects on detection in clutter. Optical Engineering}, 32(7):1675--1684, July 1993.

Depth Perception Applied to Search and Target Acquisition

Wendell R. Watkins
U.S. Army Research Laboratory
AMSRL-SL-EA
White Sands Missile Range, NM, USA 88002-5513
Tel: (505) 678-4313, Fax: (505) 678-8822, E-mail: wwatkins@arl.mil

LeRoy Alaways
U.S. Military Academy
Department of Civil & Mechanical Engineering
Mahan Hall
West Point, NY 10996
Tel: (914) 938-5517, Fax: (914) 446-7709, E-mail: il9693@usma.edu

1. SUMMARY

A search and target acquisition test was performed under an exchange scientist program with the TNO Human Factors Research Institute at Soesterberg, The Netherlands in September 1998. The test was performed at a military training base using several of the scientists from TNO wearing Dutch forest camouflage uniforms.

Sets of wide baseline stereo photos were obtained for targeted and non-targeted scenes at two sites. The targeted and non-targeted scene photos were taken on the same day within a few minutes of each other. The imagery obtained was taken with a 35 mm camera with a 200 mm lens for target ranges from 100 m to 1 km. A single field of view was used for all of the targeted and non-targeted scenes at each site. The photos were taken with color slide film and were digitized to 3K by 2K pixel resolution. These imagery data sets were used to perform search and target acquisition tests.

Preliminary analysis of single line of sight search and target acquisition observer tasks was performed for the same scenes with and without targets. Results of these observer tests are presented. Additionally, the scenes used in these tests were made into stereo pair images for observer display. There are several aspects to the display of wide baseline stereo images that must be taken into consideration for optimum depth perception for use in search and target acquisition. Rule of thumb guidelines for optimizing the depth perception of the contour of camouflaged targets versus terrain features have been derived.

Keywords: Search, target acquisition, depth perception, stereo vision, camouflage, clutter

2. INTRODUCTION

The rationale for performing this research was the results from the Distributed Interactive Systems Search & Target Acquisition Fidelity (DISSTAF) Test conducted at Fort Hunter-Liggett, CA in 1995. The visible data sets collected by the Dutch are currently being used to evaluate the camouflage, concealment, and deception (CCD) performance models for the NATO SCI-12 Working Group.[1] A group from the Army Research Laboratory collected wide baseline stereo imagery at the DISSTAF Test. The results of showing this stereo imagery to some of the observers used for the DISSTAF Test was that there are depth cues that can be used at multiple km ranges for search and target acquisition tasks.[2] These results coupled with applications of stereo vision for detecting camouflage need to be quantified for comparison with the single line of sight search and target acquisition methodology.[3] The problem of course is that there currently aren't any good models for handling clutter in imagery, even for single line of sight imagery analysis, especially when the targets are camouflaged. This deficiency was recently highlighted by James Ratches of the U. S. Night Vision at the SPIE AeroSense Symposium in an Invited Overview paper of Night Vision's efforts past, present, and future.[4] On the top of the list for future research efforts was clutter quantification.

To begin to address the issue of how to compare single line of sight search and target acquisition versus stereo vision, discussions were made between Wendell Watkins of US Army Research Laboratory (ARL) and Matthew Valeton of the Dutch Institute for Perception (TNO) at the SPIE AeroSense Symposium in Orlando, FL, USA in April 1998 for a joint research project under the exchange scientist program (APEX). APEX funding was obtained in the summer of 1998; and, a research project was conducted in September 1998 at the TNO Human Factors Research Institute, Soesterberg, The Netherlands.

2.1 Test Plan

Bearing in mind that there is no standard method for comparing single line of sight (monocular or bi-ocular) versus stereo vision for various search and target acquisition tasks, a test plan was drawn up for

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

investigating how to quantify scene clutter for both single and stereo lines of sight. The objective was to collect and analyze a database of images at a suitable test site with suitable targets for the derivation of a clutter quantification algorithm. The simplest targets to use were humans with suitable attire to match the surroundings sufficiently that all of the targets were not obvious and sufficient clutter was present to assess target placement in different clutter regions. The imagery database also had to have several lines of sight for assessment of stereo vision for comparison with bi-ocular vision performance for the same task. The human inter-ocular separation for maximum unaided depth perception ranges is about 10mrad. Several multiples of this separation distance were utilized for assessing the performance of stereo versus bi-ocular vision for the same vision task. Of course there were several more pages of details in the test plan of how to set up this field experiment and how to analyze the results.

2.2 What Was Really Done

The first issue that had to be addressed was what cameras were available for collection of the stereo imagery database. Sufficient 35 mm cameras and 200 mm lenses were obtained to set up four stereo cameras. The targets used were humans wearing Dutch forest camouflage uniforms. The test was conducted over a two-day period at the Soesterberg artillery facility where two sites were used. The first had shorter ranges (110 m to 675 m) with sunny/partly cloudy conditions. The second had longer ranges (400 to 900 m) with cloudy/rainy conditions. The first site had four camera positions with 6 m separation, and the second had three camera positions with 10 m separation. The camouflaged human targets were arrayed in sets of four for each of six different target locations. Slide photos of designated target positions, targeted scenes, and non-targeted scenes were taken at each of two test sites. This resulted in an imagery database that has 24 targets for four stereo lines of sight for the first site and three stereo lines of sight for the second site. Because photos were taken with and without the targets present, the impact of target placement and background clutter levels can be analyzed.

3. FIELD EXPERIMENTS

3.1 Measurements

With 35 mm cameras a camouflaged human can only be detected in digitized photographic film slides to a range of about 300 m. Hence, 200 mm lenses were used that yielded a field of view of 15° by 10°. To determine the positions for the targets a 35 mm camera with the 200 mm lens attached was used for viewing each of the two sites. The camera's line of sight was positioned with a conspicuous feature in the center of the field of view. A total of 24 target locations were identified for each of the two sites that represented easy to difficult targets for detection. These locations were referenced to several prominent scene features that were ranged with a binocular range finder.

Since there were only five people available for these tests, four were used as targets and one for positioning the targets and taking the photos. Hence, the targets had to be positioned in six different locations with the overall target positions ranging from about 110 m to 660 m for the first site and 400 m to 900 m for the second site. Hand-held radios were used to direct the camouflaged human targets into the correct positions. The cameras were placed on tripod mounts in a straight line that was perpendicular to the line of sight to the middle of the target scene field of view about 1.5 m above the ground. When the four targets were in their first position, the lines of sight from each of the stereo cameras to each target had to be checked to insure that the line of sight was not blocked. Then the targets held up large white cards to designate their position and one photographic slide was taken as quickly as possible from each of the stereo cameras. The targets were then instructed to turn around and hide their card and take either standing or crouching positions. By facing away the targets do not expose face or hand features that are strong detection cues for visible images. Two slide photos were taken of these targeted scenes from each of the stereo cameras. Then the targets were instructed to hide, and two slide photos were taken of these non-targeted scenes. This process was repeated six times to get the 24 target positions. The target scene for the first site without targets is shown in Figure 1. A composite target scene for the first site with all 24 targets with their white signs is shown in Figure 2.

3.2 Photo Processing

The collection of the photographic slide images for the first site took one day with sunny to partly cloudy conditions. The collection of the photographic slide images for the second site was accomplished on the next day with cloudy to rainy conditions. All of the film was then developed and digitized to 3,072 by 2,048 pixel resolution. Because of the significant changes of visibility with the rainy conditions present in the second day's testing, the first day was used for the initial analysis.

The imagery collected at the first site was collected with four different cameras. Of the four camera positions the photos from the right and left cameras were closest in terms of color matching. The center left was a little lighter, and the center-right was much lighter and more yellow even though all the cameras were set to the same exposure and aperture settings. There must have been a significant difference in the optics of the 200-mm lenses used. The color

differences caused a significant slow down in the processing of the stereo pair images with Photoshop. To begin processing the right line of sight was used as the reference. A composite picture of all of the target locations (Fig. 2) was produced by splicing all of the target photos with white location cards displayed onto the photo with the first four target positions. A display grid was placed onto this composite photo to determine an optimum size for the initial display field of view size (the trick here was not to cut targets into pieces with the edges of the individual rectangular sectors). An array of four rows of seven sectors each allows a random distribution of unshared targets into the 28 sectors. Each sector was 396 pixels wide and 264 pixel high. With the 200 mm camera lens used each sector represents a 1.9° by 1.3° field of view.

The labeling of the sectors was alphabetically for the rows, A through D (top to bottom), and numerically for the columns, 1 through 7 (left to right). The sector B4 has a bush in the center that was the conspicuous feature in the center of the camera's field of view. The non-targeted and targeted B4 sectors are shown in Figs. 3 and 4, respectively. In order to produce the targeted image in Fig. 4, portions of two different targeted images were spliced together because there were one or more of the four targets in each targeted image present in the particular sector. This type of splicing had to be performed only for a few cases. For reference purposes, the D sectors had images with terrain ranges from about 130 m to 180 m; the C sectors, from 180 m to 340 m; the B sectors, from 340 m to 520 m, and the A sectors, from 520 m to 675 m. The A sectors have truncation of the range because of the basically vertical wall produced by the tree line beyond the road at about 625 m. Also, the A sector scenes on the right have a tree line at 340 m as the lower portion of the image.

The set of right line of sight targeted sectors had 15 sectors with one target, three sectors with two targets, one sector with three targets, and nine sectors with no targets.

3.2 Image Display

In order to present the images to observers the only means available was a computer monitor display. Photoshop was used to produce sets of targeted and non-targeted sector bmp files of 792 by 528 pixels or 1.2Mbytes for the RGB color image from the original 396 by 264 pixel images. There were 56 total images for the right line of sight. In order to obtain the correct stereo image for the other lines of sight, the center terrain feature of the right line of sight was found in the other lines of sight whole scene images and a 396 by 264 rectangular image sector was cut out around this center feature. As the angular separation increased there were a few sectors that could not be matched. A random ordering of the targeted and non-targeted sectors was performed such that the ranges were mixed and targeted and non-targeted images were mixed with the additional constraint that the same sector targeted and non-targeted scenes were separated by several intervening different sector images. Finally, because of the limited number of sectors a targeted sector with an easily detected target, A4, was shown first as a learning image. Power Point was used to produce four separate slide shows of 128 scenes. The targeted and non-targeted scenes were separated by a black numbered scene with the first scene in the slide show as a black numbered scene. At present, an observer database has only been collected on the right line of sight imagery as viewed with both eyes looking at a single monitor.

4. DATABASE

4.1 Search and Target Acquisition Task

Some of the most useful search and target acquisition information can be obtained using eye tracking of the observer. Unfortunately, this type of analysis tool was not available. Hence, search time was picked as the quantifying parameter for the task of locating targets within the displayed scene. The observers were seated 1m from a computer monitor and briefed on the search and target acquisition task to be performed. The room was then darkened, and the observers were asked questions for about three minutes to allow them to become adjusted to the light level. The observers were then shown the slide show of targeted and non-targeted scenes as 20 cm wide by 13 cm high images with black borders. A stopwatch was used to measure the viewing time. The times and notes related to the location of targets or false targets found were recorded after each terrain scene was replaced with the black numbered scenes.

The observers were told that this is a search test focusing on how search times are influenced by scene content. What is desired from the observers is a concentrated effort to locate camouflaged personnel in a variety of backgrounds as quickly, yet as accurately, as possible. With this as a goal, the following are guidelines to the observer search task. (1) The targets are personnel with forest camouflage suits. (2) The personnel are either standing or crouched on the ground. (3) The targets are not perched in trees or minimally exposed with, for example, only an arm showing. (4) The personnel do not expose obvious high contrast features such as a face or hands. (5) The scenes vary dramatically in terms of range and background feature content. (6) In each scene there may be NONE, ONE, or MORE THAN ONE camouflaged personnel targets. (7) The target scenes will be alternated with black numbered scenes. (8) When a scene is shown, the task is to locate all the camouflaged personnel targets in the scene as quickly as possible. (9) Once all the targets are located the observer is to say "STOP," and the scene will be changed to the non-target display. (10) In cases where

there could be confusion the observer will be asked to point out where in the scene they saw a target or targets to determine where the target was seen. (11) As scoring criteria, the observer will be given one point for each correctly identified target, minus one point for each false target identified, and minus two points for each target missed. (12) When the observer is ready for the next scene, they are to say "READY." (13) The observer will maintain a 1-m viewing distance from the monitor display. (14) For the sake of comparison, a typical fast response time for searching a scene is one to two seconds. (15) A typical slow response time for searching a scene is around 15 seconds.

To check on whether a target or false target was identified, the observers were asked to identify where they saw the target or targets in a three by three grid within the scene. The locations are top left, top center, top right, left center, center, right center, bottom left, bottom center, and bottom right. A sample of 30 observers was used with widely varying backgrounds.

4.2 Search times and target identifications

The observer testing resulted in a 56 by 30 array of detection times and a 56 by 30 by 9 (sector sub elements) array of target or false target detection locations. The detection location data shown in Table 1 will be considered first. Table 1 gives the targets (positive identification, PI in bold numbers) and false targets (false target, FT in standard numbers) as distributed within the sector sub elements. The NULL values represent the number of sectors for which no targets or false targets were found. With a sample size of 30 observers, a difficulty factor (D) was assigned as zero for a NULL value of zero no-detections, one for one to three, and one more for every three thereafter. Hence for the 28 to 30 no-detection level the difficulty was ten or D10. The difficulty factor was also applied to the PI values. In this case if there were a PI of 30 target detections the difficulty factor (D) of zero was applied. Here for every three less in the value of PI the difficulty was increased by one. Hence, for a PI of zero to two the difficulty was ten or D10. In the targeted sector A1 the NULL value was 12 and the target in the bottom center element also had a PI value of 12. Hence the scene had a difficulty of D4 whereas the target had a positive detection difficulty of D6. It was easier to find a target in this scene because another sector element (center) had a false target with an effective PI of difficulty D6. For the case of the untargeted sector A1 the NULL value of 16 represents a difficulty for finding a target or false target of D6 comparable to the false target.

The average search time in seconds for each sector will now be addressed with respect to the NULL difficulty for both non-targeted and targeted sectors. These results are shown in Table 2. The times given represent times taken for targeted and non-targeted sectors. For the cases where there was no target present, the same sector was shown twice (9 sectors). In these cases the times for the two cases were averaged and listed in the non-target sector times.

5. RESULTS

In general the times for the non-target sectors are longer than the target sectors. In fact there are only two cases where the target sectors (A1 and B5) have longer times than the overall average search time of 6.25 seconds. This makes sense though because these were the most difficult sectors to find the targets in with difficulties of D4 and D5 respectively. If the difficulties of D0 to D3 are considered low; D4 to D7, medium; and D8 to D10, high. The average search time for the non-targeted sectors with medium difficulty is 6.7 s and for high difficulty 7.5 s. For the targeted sectors with medium difficulty the time is 7.5 s and for low difficulty only 4.0 s. In essence if a target real or false is easy to detect, the search time drops significantly.

The range of average search times for individual observers was from 1.75 s to 13.88 s. There was a correlation between poor overall scores and longer search times. This is presented in Table 3. To better compare the results of the different observers with respect to the differences between times taken to search individual sectors, the times for each observer were normalized by dividing each time by their individual average search time. When this was done there were 852 sectors where no target or false target was detected taking an average normalized time of 1.15. There were 828 sectors where targets or false targets were found taking an average normalized time of 0.81. To better see how much time is taken to find a target or false target, the average normalized detection and no-detection times for each sector are compared versus the NULL difficulty in Table 4. Column 2 in Table 4 gives the normalized times to detect a target or false target. As the scene becomes more difficult (i.e., the NULL value increases, and there are fewer of the 30 observers who detect targets or false targets) the normalized time taken becomes longer. Also, in general it takes longer to determine that there is no target or false target present than when there is. When the false targets present are very target-like as in the case of most of the D4 and D5 samples, the detection time, TG-TIME, is short and the non-detection time, ND-TIME, is long. In addition, the number of false targets versus real targets detected increases as the NULL value increases. This is similar to over training a neural net.

There were a few examples of how moderate to difficult targets are missed in scenes when there is an easier target or false target detected first. The sector B4 had three targets present with PI difficulties of D10, D6, and D1 located, respectively, in the left center, right center, and center of the sector. This

image gave a good example of how the human detection process works. When a search and target acquisition task is given, a fuzzy notion of what the target of interest is is formulated. The presented scenes are searched for this fuzzy target. If a detection of a real or false target is made, the target construct becomes well defined and the scene search is rapidly completed thereafter even if multiple targets are present and detected. This refinement in the target sought can cause targets to be missed. In sector B4 there is a fairly easy target to detect right in the middle. This is a standing target. The crouching target to the left and away from the tree line was only detected if it was seen first. Only two of the 30 observers accomplished this. Both of these observers were able to then detect the center easy target but did not detect the moderately difficult target on the right center. A similar occurrence happened in sector A1 where there was a bush that very much resembled a standing target in the center of the sector. This made the detection of the crouching real target in the bottom center more difficult.

6. STEREO VISION

The initial viewing of the stereo slide shows revealed some distinct problems. The images in the closest sectors (C and D) could not be fused for the field of view of the entire sector. There simply was too much parallax. At 110 m the approximate 1.9 m high human targets represent about 90% of the sector image height or 238 pixels. At 650 m the human targets represent only about 15% of the sector image height or 40 pixels. With a 6 m platform separation between the right and right-center cameras the resulting shift between the bottom and top elements of the scenes in the D sector is 1.8 m or 225 pixels with the standing target experiencing 90% of this shift from bottom to top. In the C sector the parallax shift bottom to top is 5.0 m or 180 pixels. This time a 1.9 m target in the bottom of the scene would represent only 45% of the height with only about 81-pixel parallax shift from the bottom to top of the target. Stereo fusion at this range was possible but not comfortable. Finally, in the B sector the parallax shift bottom to top is 6.3 m or 145 pixels. Now, the 1.9 m target in the bottom of the scene represents just 25% of the height with only about 36 pixel parallax shift from the bottom to top of the target. These images could be easily fused and showed good depth perception.
Hence, to be able to compare the results of single line of sight to stereo vision for the near targets would require a display of field of view about one third the one that was used for the closest sectors.

7. CONCLUSIONS

A search and target acquisition test was conducted that provided single and wide baseline stereo imagery for observer testing. The database contains the same scene with and without camouflaged human targets present. The analysis of imagery from the first of two sites has resulted in several interesting findings. First, a simple search task showed that search times are significantly longer for scenes where no target or false target is detected. Second, there was little difference in total search time for one or many detected targets. Third, as the normalized time that a scene is viewed increases the probability of false target detection also increases. Finally, the use of stereo vision for reducing the clutter level in search and target acquisition tasks has promise, but requires care in assessing. It cannot be done for short range targets without using multiple fields of view.

8. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Matthew Valeton and the rest of the Vision and Imaging research team of the TNO Human Factors Research Institute, Soesterberg, The Netherlands. Without their assistance in assembling the equipment, in processing the photographic slides, and in providing the targets and test range needed for this experiment, this research could not have been possible. The human target especially will remember tromping through heather while getting soaked by rain and trying to guess how to respond to unintelligible hand-held radio messages.

9. REFERENCES

[1] Bijl, P., Kooi, F.L. and Valeton, J.M., *Visual search performance for realistic imagery from the DISSTAF field trials* (Report TM-97-A055), TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1997.

[2] Watkins, W.R., *Multispectral Image Processing: The Nature Factor*, Proc. of the SPIE, Vol. 3545, October 1998.

[3] Watkins, W.R., Jordan, J.B., and Trivedi, M.M., *Novel applications of hypersterio vision*, Proc. of the SPIE, Vol. 3310, March 1997.

[4] Ratches, J.A., *Night vision modeling: historical perspective*, Proc. of the SPIE, Vol. 3701, April 1999.

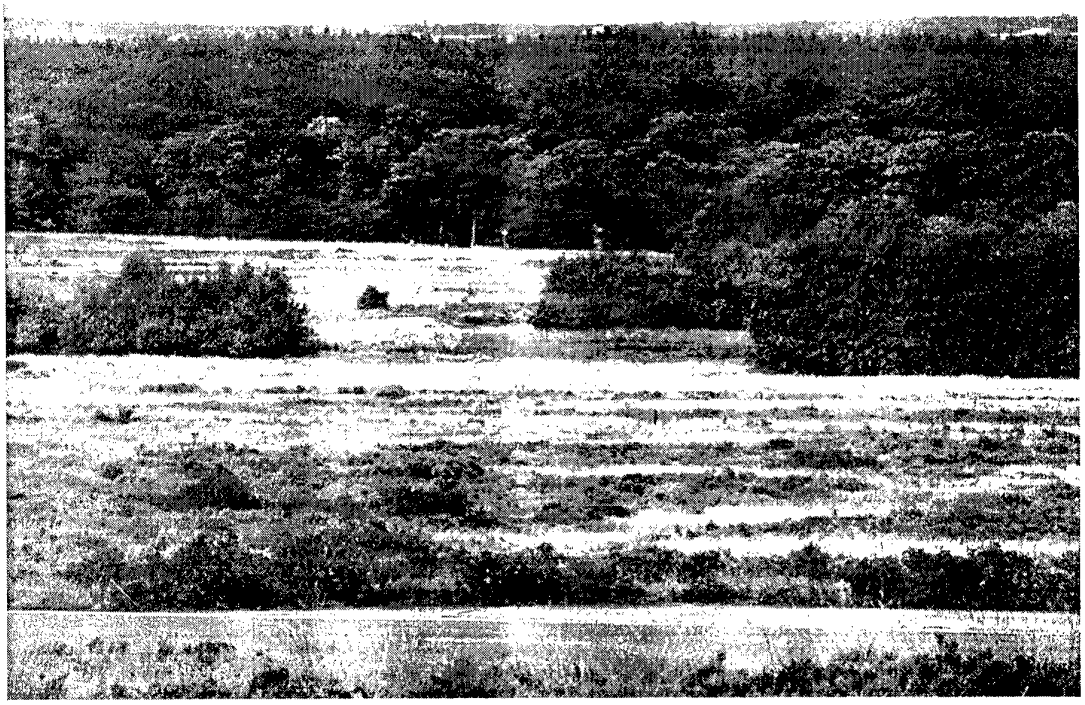


FIGURE 1. Whole scene with no targets.

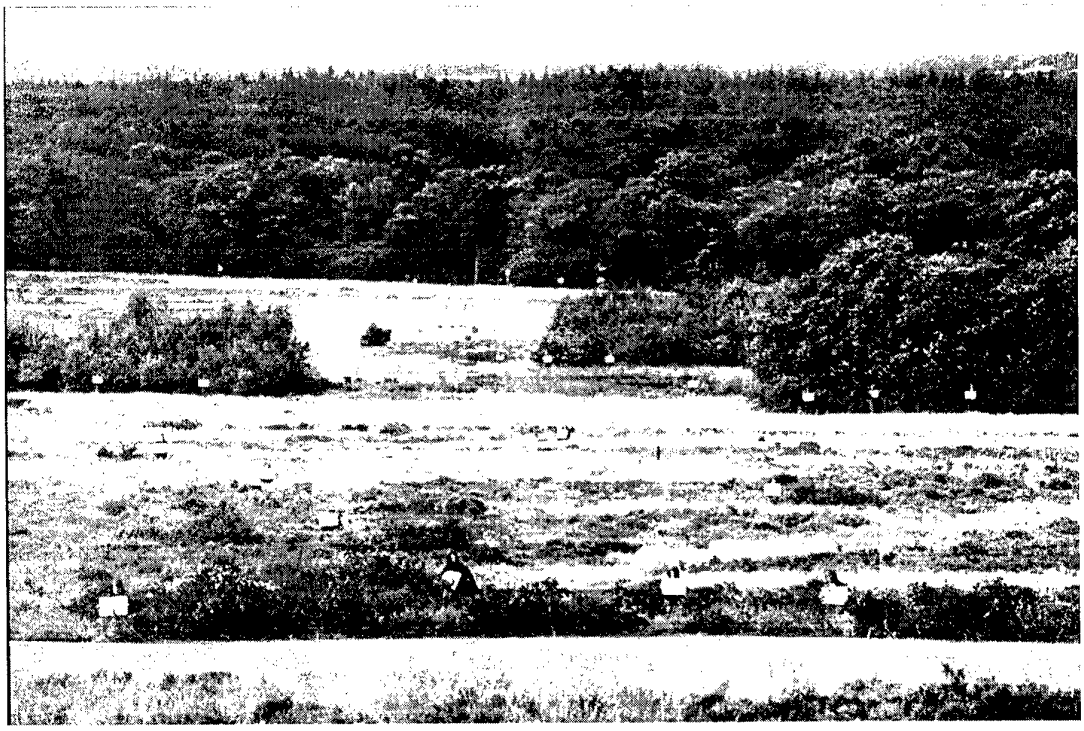


FIGURE 2. Whole scene with target positions designated with large white cards.

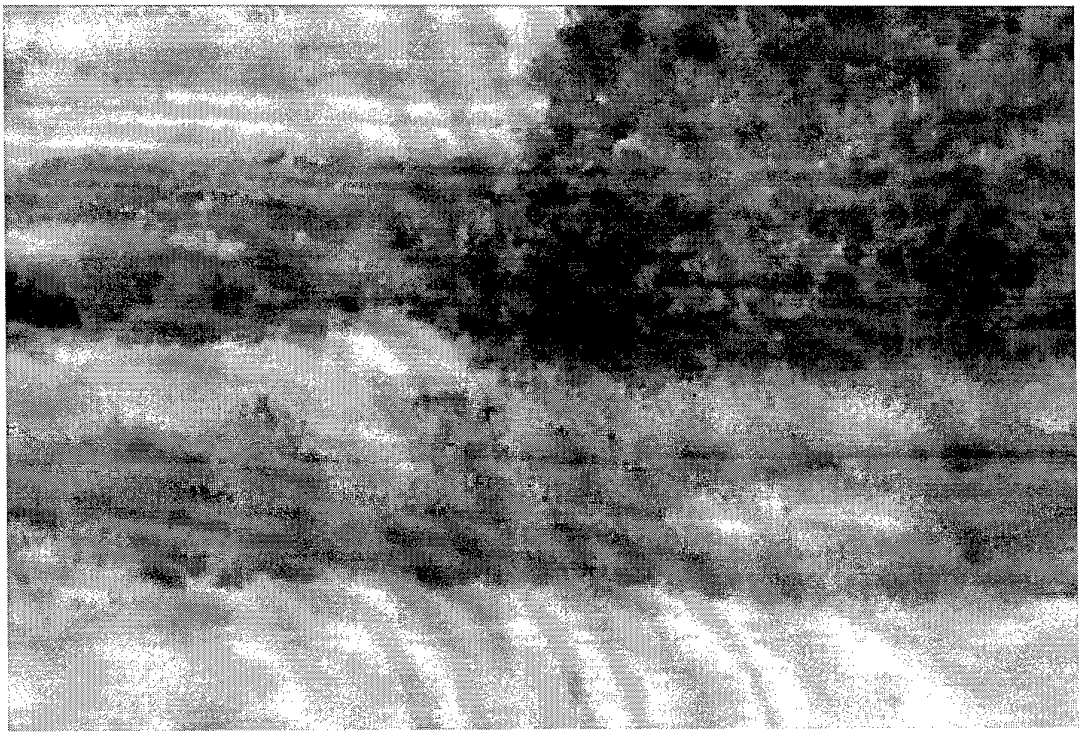


FIGURE 3. Non-targeted sector B4.

FIGURE 4. Targeted sector B4.

TABLE 1. SECTOR LOCATIONS OF TARGET AND FALSE TARGET DETECTIONS

A1	PI 12	0	0	0	A1	PI 0	0	0	0
TG (1)	FT 16	0	11	3	BK	FT 20	4	12	2
#22	NULL 12	1	**12**	1	#6	NULL 16	1	0	1
A2	PI 28	0	0	0	A2	PI 0	1	0	0
TG (1)	FT 3	**28**	3	0	BK	FT 10	4	3	2
#53	NULL 2	0	0	0	#38	NULL 24	0	0	0
A3	PI 0	0	0	0	A3	PI 0	0	0	0
TG (0)	FT 19	0	12	7	BK	FT 12	0	7	5
#28	NULL 20	0	0	0	#15	NULL 24	0	0	0
A4	PI 30	0	0	0	A4	PI 0	0	0	0
TG (1)	FT 3	3	**30**	0	BK	FT 13	10	1	1
#1	NULL 0	0	0	0	#48	NULL 17	1	0	0
A5	PI 0	0	0	0	A5	PI 0	0	0	0
TG (0)	FT 4	2	0	1	BK	FT 7	4	1	0
#43	NULL 27	0	1	0	#18	NULL 24	1	1	0
A6	PI 0	0	0	0	A6	PI 0	0	0	0
TG (0)	FT 39	34	5	0	BK	FT 40	38	2	0
#16	NULL 12	0	0	0	#50	NULL 10	0	0	0
A7	PI 0	0	0	1	A7	PI 0	0	3	1
TG (0)	FT 3	0	1	1	BK	FT 6	0	2	0
#47	NULL 28	0	0	0	#27	NULL 25	0	0	0
B1	PI 30	0	0	0	B1	PI 0	0	0	1
TG (1)	FT 2	30 + 2	0	0	BK	FT 6	1	3	0
#41	NULL 0	0	0	0	#26	NULL 25	0	1	0
B2	PI 27	0	0	0	B2	PI 0	0	0	0
TG (1)	FT 2	**27**	1	1	BK	FT 7	1	2	4
#23	NULL 3	0	0	0	#3	NULL 23	0	0	0
B3	PI 30,23	0	**23**	0	B3	PI 0	0	0	1
TG (2)	FT 0	**30**	0	0	BK	FT 3	0	1	1
#55	NULL 0	0	0	0	#36	NULL 27	0	0	0
B4	PI 2,29,14	0	0	0	B4	PI 0	0	0	0
TG (3)	FT 0	**2**	**29**	**14**	BK	FT 4	2	2	0
#29	NULL 0	0	0	0	#14	NULL 27	0	0	0
B5	PI 12	0	0	0	B5	PI 0	0	0	0
TG (1)	FT 6	4	1	0	BK	FT 1	0	1	0
#4	NULL 15	0	**12**	1	#45	NULL 29	0	0	0

TABLE 1
CONT.

B6	PI 30,30	0	0	0	B6	PI 0	1	0	0
TG (2)	FT 1	0	0	0	BK	FT 8	0	1	0
#35	NULL 0	0	30+1	30	#24	NULL 25	0	4	2
B7	PI 30	0	0	0	B7	PI 0	0	0	2
TG (1)	FT 4	2	0	0	BK	FT 17	3	0	2
#10	NULL 0	1	30	1	#56	NULL 17	7	1	2
C1	PI 30	0	0	0	C1	PI 0	0	0	0
TG (1)	FT 6	2	4	30	BK	FT 11	5	3	0
#11	NULL 0	0	0	0	#54	NULL 25	0	2	1
C2	PI 30	0	0	0	C2	PI 0	0	0	8
TG (1)	FT 2	0	30	0	BK	FT 15	1	0	0
#43	NULL 0	1	1	0	#32	NULL 20	5	0	1
C3	PI 0	2	2	2	C3	PI 0	2	0	7
TG (0)	FT 7	0	0	0	BK	FT 12	0	1	2
#17	NULL 25	1	0	0	#2	NULL 20	0	0	0
C4	PI 29	0	29	1	C4	PI 0	3	0	1
TG (1)	FT 2	0	0	0	BK	FT 9	0	1	0
#49	NULL 1	0	0	1	#33	NULL 24	3	0	1
C5	PI 0	3	5	3	C5	PI 0	0	4	0
TG (0)	FT 19	1	0	4	BK	FT 14	1	2	5
#31	NULL 17	0	0	3	#12	NULL 20	0	0	2
C6	PI 30,28	2	0	0	C6	PI 0	5	1	2
TG (2)	FT 9	30 + 5	1	1	BK	FT 17	2	1	4
#5	NULL 0	28	0	0	#44	NULL 20	1	0	1
C7	PI 0	0	0	1	C7	PI 0	0	0	0
TG (0)	FT 10	1	4	0	BK	FT 10	1	3	3
#37	NULL 24	2	1	1	#21	NULL 24	1	0	2
D1	PI 30	0	1	0	D1	PI 0	1	0	2
TG (1)	FT 2	0	1	0	BK	FT 6	0	1	0
#40	NULL 0	0	30	0	#20	NULL 25	0	1	1
D2	PI 30	0	0	30	D2	PI 0	0	0	0
TG (1)	FT 4	0	1	0	BK	FT 17	2	2	0
#13	NULL 0	2	0	1	#51	NULL 18	6	0	7
D3	PI 30	0	1	0	D3	PI 0	1	0	2
TG (1)	FT 8	1	3	2	BK	FT 17	3	1	5
#46	NULL 0	1	0	30	#30	NULL 19	5	0	0

TABLE 1
CONT.

D4	PI 0	2	1	0	D4	PI 0	1	3	1
TG (0)	FT 9	1	0	0	BK	FT 12	2	1	1
#19	NULL 22	1	4	0	#8	NULL 24	0	2	1
D5	PI 30	1	0	0	D5	PI 0	2	0	0
TG (1)	FT 7	0	2	0	BK	FT 11	0	3	3
#52	NULL 0	0	**30**	4	#39	NULL 24	0	1	2
D6	PI 30	1	0	0	D6	PI 0	4	0	2
TG (1)	FT 4	1	1	1	BK	FT 18	0	6	0
#25	NULL 0	0	**30**	0	#9	NULL 21	1	4	1
D7	PI 0	0	0	1	D7	PI 0	0	0	0
TG (0)	FT 12	1	6	1	BK	FT 5	1	4	0
#7	NULL 20	3	0	0	#42	NULL 26	0	0	0

TABLE 2. SEARCH TIMES FOR TARGETED AND NON-TARGETED SECTORS

		1	2	3	4	5	6	7
A	no target	7.2 / D6	8.7 / D8	7.9 / D8	7.5 / D6	9.7 / D9	6.1 / D4	6.4 / D9
	target	**7.4 / D4**	**5.2 / D1**	**NT**	**3.7 / D0**	**NT**	**NT**	**NT**
B	no target	6.2 / D9	6.9 / D8	7.6 / D9	6.5 / D9	7.7 / D10	8.7 / D9	6.2 / D6
	target	**3.7 / D0**	**4.3 / D1**	**4.0 / D0**	**4.6 / D1**	**7.5 / D5**	**3.4 / D0**	**3.8 / D0**
C	no target	5.6 / D9	6.7 / D7	6.5 / D8	7.4 / D8	5.9 / D7	6.4 / D7	7.7 / D8
	target	**3.1 / D0**	**3.4 / D0**	**NT**	**3.3 / D1**	**NT**	**4.5 / D0**	**NT**
D	no target	7.4 / D8	6.9 / D6	7.4 / D7	8.1 / D8	8.6 / D8	8.0 / D7	7.2 / D8
	target	**3.9 / D0**	**4.6 / D0**	**4.1 / D0**	**NT**	**3.5 / D0**	**4.1 / D0**	**NT**

TABLE 3. OBSERVER SCORES AND TIME RANKING

PERSON	SCORE	TIME(s)	T-RANK	5R-AVG
1	18	3.88	8	----
2	17	2.16	2	
3	17	7.63	21	14.4
4	17	11.96	29	
5	16	4.35	12	----
6	15	13.88	30	----
7	14	3.32	4	
8	14	3.51	5	11.6
9	13	5.96	17	
10	10	2.59	3	----
11	10	1.75	1	----
12	9	3.73	7	
13	8	4.78	14	10.2
14	8	4.29	11	
15	7	7.71	23	----
16	7	4.17	10	----
17	7	6.76	18	
18	3	4.85	15	14.2
19	3	4.57	13	
20	2	4.08	9	----
21	-1	7.75	24	----
22	-7	5.68	16	
23	-12	9.6	25	19
24	-14	3.63	6	
25	-14	7.56	20	----
26	-14	6.98	19	----
27	-16	11.98	28	
28	-25	9.69	26	23.6
29	-26	11.04	27	
30	-95	7.71	22	----

TABLE 4. NORMALIZED TIMES AS A FUNCTION OF TARGET DIFFICULTY

D	TD-TIME	#	ND-TIME	#
0	0.64	13	-----	0
1	0.74	4	0.8	4
2	-----	0	-----	0
3	-----	0	-----	0
4	0.9	3	1.29	3
5	1.1	1	1.48	1
6	1.05	5	1.13	5
7	1.07	8	1.19	8
8	1.34	10	1.24	10
9	1.12	10	1.08	10
10	0.96	2	1.11	2

METHODS FOR DERIVING OPTIMUM COLOURS FOR CAMOUFLAGE PATTERNS

K.D. Mitchell, C.R. Staples
Science and Technology Division.
Defence Clothing and Textiles Agency
Flagstaff Rd
Colchester
Essex
CO2 7SS
United Kingdom
E-mail: kdmitchell@dcta.demon.co.uk

1. SUMMARY

The majority of camouflage patterns have been designed subjectively with only the colour aspect conforming to certain constraints such as average colour and luminance. Given the power of modern computing it should be possible to design scenario specific camouflage from calibrated colour imagery. The Defence Clothing and Textiles Agency is at present working on such a system. This capability will allow us to design and test patterns in a digital environment before field trials are carried out. This system will allow us to design patterns for specific scenarios such as coniferous treelines, deciduous treelines, summer, winter etc. It should also lead to highly effective patterns, as early validation can be carried out using a target detection model followed by photosimulation using a digital implantation technique. Once validated in the digital environment, a field trial using live observers can be carried out.

In the design of a pattern, there are two major factors to take into account: the multi-level structure of a background and the many colours present. A method of designing scenario specific patterns needs to reduce the many hundreds of colours to a workable number of colour centres, usually between three and six. There is also the need to assess the structure present and produce a structure for the pattern, which should be multi-level to allow the pattern to be effective at various ranges.

In this paper, we will review the results obtained from the initial study on reduction of the number of colours and colour centre choice.

Keywords: Colour choice, patterning, optimisation routine

2. INTRODUCTION: TRADITIONAL METHODS OF CAMOUFLAGE DESIGN

Traditionally methods of camouflage design for materials have been mainly subjective with the only constraints being the colours used and the average luminance of the overall pattern.

The methods of traditional design involve the designer viewing a background and using their skill and judgement to devise a pattern which will be effective. This pattern must then be trialled to assess its effectiveness and may also undergo further validation techniques such as photosimulation. Any validation routines have by necessity to take place over various scenarios and compare several camouflage schemes. The personnel and time needed for such validation makes the costs very high. The pattern designed often has to be applicable to several theatres of operation i.e. temperate zones, jungle environments, arctic and desert and must be a good average to account for the diversity within each background.

3. THE GENERATION OF PATTERNS USING A COMPUTER BASED METHODOLOGY

The generation of patterns from digital images using computers gives the capability to design scenario specific patterns, relatively quickly and cheaply. There are three discrete parts to the design of a new camouflage pattern using a digital methodology.

1. A method of texture analysis and generation
2. A method of optimising the choice of colours from those found in a background so the pattern is most effective either against a specific background or over a wide range of scenarios.
3. A target detection model which will allow us to measure the relative effectiveness of camouflage schemes

Parts 1 and 3 can be carried out using either colour or monotone images but, for an effective visual camouflage part 2 is a highly important factor.

A methodology which allows us to carry out textural analysis could also be used to design a pattern which has first and second orders statistics that resemble those of the background. The in-service U.K. pattern has an average colour which resembles an average colour of a set number of treelines. The design only incorporates the first order statistics of a series of backgrounds. Second order statistics are used to describe the textural elements of the particular region being analysed. This ability to design the first and second order statistics of the pattern and the use of target detection models allows us to predict the relative effectiveness of several patterns in a digital environment. This reduces the initial costs, as we do not have to go through such a large-scale trial and do not have the expense of making life-size uniforms or designs for vehicles. The use of colours in a pattern can determine its effectiveness. A method of optimising the limited number of colours used is highly desirable.

4. HOW ARE COLOURS USED IN A CAMOUFLAGE SCHEME?

The colours in a camouflage pattern or scheme are ideally used to allow the target to blend in to the background. This is done on two levels. Firstly, the colours used are those found in the background. For rural camouflage, these are browns, greens and black resembling those found in a natural scene. Secondly, the colours form a pattern which, it is hoped match that of the background and reduces any visual cue given by the outline. It may be said that the colours and textured pattern used in a camouflage scheme are equally important from a detection point of view. A good pattern's effectiveness will be reduced by bad colour choice and good colours will be ineffective if the patterning is poor. It should be noted that no matter how good the colours or pattern at very long ranges

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

both are inconsequential e.g. at long ranges where the background appears monochromatic and atmospheric effects dominate[1]. At closer ranges, the better the patterning choice and colours used, the shorter the detection range. However, for vehicles in particular, there are ranges where the vehicle just cannot be disguised.

5. COLOUR CHOICE

The human eye can see all the colours in a specific background but has a problem if asked to reduce these colours to a given number for a best fit. The human eye tends to blend the colours it sees where as a digital image taken at close range will only average over a very small area, depending on how the digitisation is done. In addition, humans tend to have a bad visual memory for exact colours, whereas a calibrated digital image will contain exact data for a specific scene at a given moment in time. As a result, choosing the colours to be used to optimise the effectiveness of a pattern is a task more suited to digital calculation than to human judgement.

6. COLOUR REDUCTION/OPTIMISATION ROUTINE

For our colour reduction/optimisation routine, we decided on a methodology which concentrated on a particular Region Of Interest (ROI). The values which describe the colour of each pixel, in a 3D colour space, are run through a mathematical routine which finds the best fit colour centres for the colour population of the ROI.

Before using the routine, decisions have to be made as to how its various capabilities are going to be utilised. It is necessary to decide how many colour centres are to be used, and whether a number of those centres are to be predetermined or all are to be optimised.

The first step in the use of the actual program is the section of a ROI from the image. If we use the whole image, which might be up to 4000x4000 pixels, the length of time needed to run the routine may extend into a number of weeks. A good size for a ROI is up to 200x200 pixels (although this will have to be run overnight if a large number of colour centres is to be used). The size of the ROI is up to the individual user, but it should contain a good cross-section of the colours found in the background as well as some of the textural elements. Figure 1 shows a region which is 100x100 pixels in size and contains good information on the type of background we want to be camouflaged against.

Fig 1: ROI taken from original image 100x100 pixels

As stated, once the region of interest has been chosen, the routine then converts the RGB values for each pixel to the Lab values. The conversion to Lab colour space allows us to describe the colours in a colourspace which resembles how the colours are actually perceived by humans[2]. This conversion is described in more detail in Houlbrook[3]. Once the conversion has taken place the values are plotted as in Figure 2. This plot allows the operator to view the most populated volumes and so place the initial colour centres near these population centres. This allows the routine to run quicker and ensures that in the later stages all of the colour centres are interrogated.

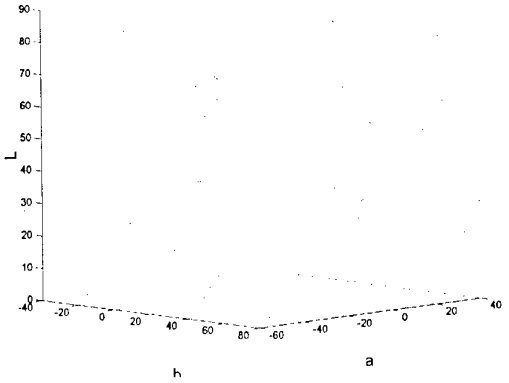


Fig 2: Plot of the Lab values of the population of the pixels found in Fig 1

The next step of the routine is an iterative step, which ceases when the best colour centres are found. During the initial step of the iteration process, each of the pixel points interrogates the initial colour centres and assign themselves to their respective nearest colour centre. Figure 3 shows a simplification of this initial assignment process.

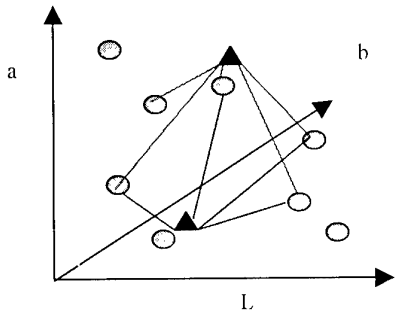


Fig 3: Showing an example of Pixel co-ordinates (●) in colourspace being assigned to colour centres 1 (▲) and 2 (▲)

Once the co-ordinates of the pixels have all been assigned to the most appropriate colour centre an average is taken of the population to find the centre point. This centre point is assigned as the new colour centre (See Figure 4)

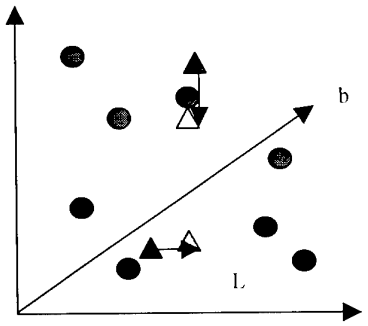


Fig 4: After step one the colour co-ordinates are assigned to their respective colour centres (● and ●) and the colour centres relocated to the average of the population (colour centre 1'▲ and colour centre 2'▲).

This stage is repeated with all of the pixel points interrogating each of the new colour centres and the averaging process repeated. The iteration process ceases when the difference between the newest colour centre and the previous colour centre is within a predefined limit. These final colour centres are then written to file for retrieval later. The final step of the routine is to change the pixel co-ordinate values to those of the colour centre to which it is assigned. So the whole population of that colour centre has the same Lab values. The routine will then show a visual representation of the original image with the new Lab values. This allows the operator to do a visual comparison as a check. Although the human visual system is not good at remembering exact colours when asked to compare two images, it can judge quite effectively if the colours chosen appear to be correct or not. Figure 5 gives an example of this phenomenon, in that although there are approximately 9000 fewer colours used than in Figure 1, the overall impression is not greatly diminished in the reduction to 10 colours.

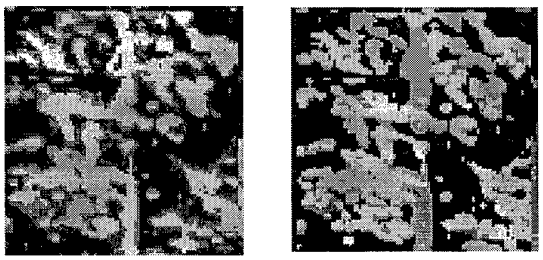
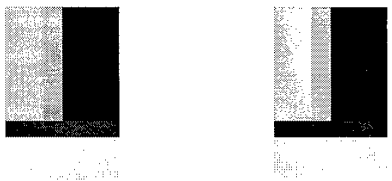


Figure. 5: Visual representation of results of the routine (i) shows optimisation to 10 colours (ii) shows optimisation to 4 colours.

7. RESULTS OBTAINED USING SYNTHESISED IMAGES

As an initial test of the routine, it was decided to use simple images constructed in Adobe PhotoShop. The rationale behind using images synthesised in PhotoShop is that images will be initially viewed using this software. Using PhotoShop we can create images with known Lab and RGB values. This allows us to compare the Lab values in PhotoShop to the values obtained from the routine when digital RGB is the only input. This gives us a good insight to how good the routine is at calculating Lab's from RGB.

The image in Figure 6 consists of 66 pixels and was created so we could carry out calculations both manually and using the routine to check that the initial conversion was correct.



(i) (ii)

Fig 6: Image constructed to check conversion from digital rgb to Lab (i) and image obtained at end of routine (ii)

	PhotoShop			Routine		
	L	a	b	L	a	b
C1	77	-34	18	77.8	-32.5	18.0
C2	63	-28	15	63.5	-27.0	15.1
C3	9	3	4	8.9	1.8	3.3
C4	0	0	0	0	0	0
C5	33	-1	27	33.0	-1.8	26.4
C6	84	-6	76	83.4	-7.8	78.0

Table 1: Comparison of the Lab's quoted in PhotoShop and those obtained as results from the routine

As can be seen from Table 1 the results obtained from the routine are similar to those quoted in PhotoShop. The differences can in part be accounted for by rounding errors.

8. RESULTS USING REAL IMAGERY

In addition to the synthesised imagery, the colour reduction routine has been successfully applied to real imagery. The results when applied to real imagery are shown in Table 1. The original image is shown in Fig 1. (The printing process has degraded these images to a certain degree.)

It may be asserted however that the colour-reduced imagery of real scenes is, by eye, a good match for the original imagery. Where we allow 10 colour centres, the scenes are barely discernible from the originals. This fidelity naturally falls with the number of allowed colour centres.

This match to the background is not, however, the purpose or test of the routine. That will come with the application of chosen colours in the creation of new camouflage schemes. These will be tested for their ability to blend into the background - by detection modelling, photo-simulation and ultimately, test in the real world

9. USE OF THE COLOUR REDUCTION ROUTINE

As mentioned there are several variations which can be carried out using the colours reduction routine. Firstly we can carry out the optimisation routine which will allow us to reduce the number of colours in an image to a given number. These colours will also be optimised to best describe the colours in the original image. Secondly, when inputting the starting colour centres we can choose to lock these centres so that the colour reduction takes place to these colour centres. This is particularly useful if the colours have been pre-determined and it is these colours you want to use. You can now see how those colours compare to those in the background. We also obtain information on the proportion of each colour in the background.

10. CONCLUSION

We have described in this paper a routine, which can be used to derive optimum colours for a camouflage pattern using calibrated digital imagery. It has been recognised that the optimisation of colours for a pattern is desirable. Optimised colours used in a pattern can reduce the ranges at which targets become visible in specific scenarios.

Colour as has been described is an intrinsic part of a camouflage pattern. In the rush to devise a digitally based method for pattern design, the use of the best colours has been largely overlooked. This work addresses this oversight and represents an important step in the development of a more complete digital pattern design tool. A routine such as this allows the optimisation of colours for a scenario without the need for extensive field trails and so cuts the time needed for the design of an effective pattern for that scenario.

11. REFERENCES

1 Phillips, P.L., "Colour Tolerances for Texture Investigation", Final BAe report on MoD Contract No. A82a/2782, 1983.

2 McDonald, R., "Colour Physics for Industry", Society of Dyers and Colourists, 1997.

3 Houlbrook, A.W., "The Development of an Image manipulation Facility for the Assessment of CCD", This Proceedings, 1999.

THE DEVELOPMENT OF AN IMAGE MANIPULATION FACILITY FOR THE ASSESSMENT OF CCD

A. W. Houlbrook
Science and Technology Division
Defence Clothing and Textiles Agency
Flagstaff Road
Colchester
CO2 7SS
United Kingdom
E-mail: awhoulbrook@dcta.demon.co.uk

1. SUMMARY

The assessment of CCD systems using photosimulation is the tried and tested alternative to performing live observer trials. The greater control over photosimulations allows an increased level of confidence in the results of any comparisons. It also requires less time in the field for a smaller number of personnel. The next step along this route would be a method that required no time in the field. Virtual reality systems, however, do not yet produce the level of realism required. An alternative, perhaps, is to place targets generated by VR software into a scene recorded photographically. Such a system would digitize a slide of a background scene in a controlled manner and allow the realistic implantation of an artificially created target. Reproduction would be achieved using a calibrated film printer. The majority of the reprinted scene would remain identical to the original slide. The methods used to enable the calibration of the equipment used, and the process of comparing information from digital rgb and Lab colour spaces are discussed in this paper.

This image manipulation facility has the potential to bypass the field trial phase of CCD assessment. It could be used to assess CCD more thoroughly by using a variety of background scenes or the same scene at different times of the year. Targets created from CAD models could be assessed. It could be used to determine the effectiveness of potential CCD measures in areas which are not readily accessible. Overall, this system adds a new level of flexibility and completeness to photosimulation.

Keywords: Assessment, photosimulation, imagery, slide, scanner, printer, colour, calibration.

2. INTRODUCTION

The assessment of CCD systems using photosimulation is the tried and tested alternative to performing live observer trials (Ashforth et al., 1991). The greater control over photosimulations allows an increased level of confidence in the results of any comparisons performed in this manner. It also requires less time in the field for a smaller number of personnel. The next step along this route would be a method that required no time in the field. Virtual reality systems would appear to offer this capability. However, they do not yet produce the level of realism required. An alternative, perhaps, is to place targets generated by VR software into a scene recorded photographically. A photosimulation could then be performed on the modified imagery. Such a system would have to digitise a slide of a background scene in a controlled manner. This would then allow the realistic implantation of an artificially created target. Reproduction of the modified scene as a photographic slide would be achieved using a calibrated film printer. In an ideal system, the majority of the reprinted scene would remain identical to the original slide.

The following describes the efforts made by the DCTA towards creating a facility to perform such image manipulation.

3. APPARATUS

The principal components of an image manipulation facility should be a computer capable of handling large image files of 100Mb or more, and a slide scanner and film printer with resolutions close to that of photographic film.

The computer used to control both the slide scanner and the film printer is a Silicon Graphics Indigo2 running the IRIX 6.2 operating system, with 384Mb of RAM. Silicon Graphics computers are designed to manipulate graphics easily, having a very capable graphics card built in as standard.

The slide scanner is a LeafScan-45. It operates by moving the slide between a filtered fluorescent light source and a linear array camera. It can operate at a variety of resolutions up to 5000 dpi. Controlling software called Image Proof is used to set up and calibrate the device, allowing the user a choice of aperture and exposure time combinations. It also incorporates a prescan facility and several adjustment controls for the image. The shadow control affects the brightness of low intensity pixels and the highlight control that of high intensity pixels. The gamma control affects the gamma of the image, in effect the linearity of the dark to light transition. The controls act on the image as a whole and each of the red, green and blue channels separately. These can be used to correct colour discrepancies due to the scanning process prior to a high-resolution final scan.

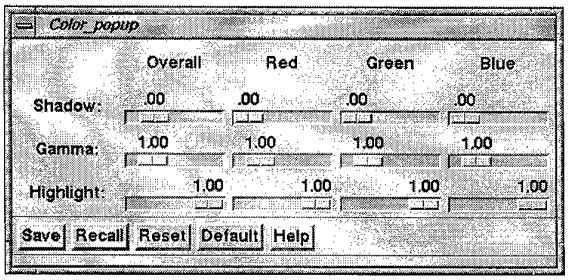


Figure 1 Scanner control window

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

The software has the capability to perform a low resolution prescan on a slide. It also allows the user to probe the prescan image for the colour information from each pixel.

The film printer is a Mirus Galleria. It prints to any type of 35mm film using a film specific look-up table to adjust its output. The look-up table contains details of the resolution required and controls for global contrast and brightness and channel specific brightness. These controls can be used to calibrate the device. It has a maximum resolution of 3333 dpi.

4. COMPARING COLOUR SPACE INFORMATION

It has been mentioned earlier in this paper that it is essential to be able to calibrate both the slide scanner and the film printer. To do this the attributes of the image must be compared before and after both the scanning and printing processes. A convenient start point for the scanning process is the colour information derived from a prescan, with the controls set to the default values, of a slide using the probe function. The aim of this operation is to scan the slide retaining the colorimetric properties of the projected image, therefore these values should be used as the target for the calibration. The film printer's start point is the scanned image, which should have the same colorimetric properties of the projected slide. Its target is the colorimetric properties of the projected image. The devices employed in this facility view and manipulate images in a 24-bit colour system, 8 bits or 256 levels of intensity for each of the red, green and blue channels. This system is derived from the signals required by the digital to analogue converters controlling the voltage to the electron guns for each channel in a colour monitor. However, the projected slide is measured using a telespectroradiometer, which produces a photometric output such as Lab values. It is therefore necessary to be able to convert digital rgb to photometric Lab and vice versa. This transform has two important stages, the transform from a digital rgb system to a photometrically based RGB system and then the further transform to Lab.

4.1. Linearisation of rgb using the gamma function

The relationship between the voltage applied to the electron gun and the luminous output of the screen is known as the gamma function.

$$Y = K(g)^{\gamma}$$

Where

Y is the display luminance /cdm^{-2}

g is the pixel grey level value ($g \in [0,255]$)

K is a proportionality constant

γ is the display gamma coefficient

Here, however, the display luminance is not measured but taken from PhotoShop's use of both rgb and Lab colour spaces, L being converted to Y assuming D65 illuminant. The maximums are 255 for rgb and 100 for Y (L is also 100). The mid points are 169 for rgb and 50 for Y (L is 76.07). Two simultaneous equations can be constructed:

$$50 = K * 169^{\gamma}$$

$$100 = K * 255^{\gamma}$$

Solving the above equations gives:

$$K = 0.008810382471$$

$$\gamma = 1.684993781.$$

These values can be put back into original equation to produce the following transforming equations for each of the red, green and blue channels:

$$R = 0.008810382471 * r^{1.684993781}$$

$$G = 0.008810382471 * g^{1.684993781}$$

$$B = 0.008810382471 * b^{1.684993781}$$

Where RGB represent the photometric values and rgb represent the digital values.

The reverses of these transforms are:

$$r = \left(\frac{R}{0.008810382471}\right)^{\frac{1}{1.684993781}}$$

$$g = \left(\frac{G}{0.008810382471}\right)^{\frac{1}{1.684993781}}$$

$$b = \left(\frac{B}{0.008810382471}\right)^{\frac{1}{1.684993781}}$$

It should be stressed that these RGB photometric values are only relevant to this system and do not directly relate to the CIE standard observer.

4.2. Transform from RGB to XYZ

To convert from these RGB values to XYZ tristimulus values, a set of transform equations are required (Hunt 1987). Such as:

$$X = A_1R + A_2G + A_3B$$

$$Y = A_4R + A_5G + A_6B$$

$$Z = A_7R + A_8G + A_9B$$

Where $A_1 - A_9$ are constants.

In order to determine these constants it is necessary to perform the following steps.

The chromaticities (a_1-a_9) of the primaries of European colour monitors are:

	x	y	z
Red	$0.64 = a_1$	$0.33 = a_2$	$0.01 = a_3$
Green	$0.29 = a_4$	$0.60 = a_5$	$0.11 = a_6$
Blue	$0.15 = a_7$	$0.06 = a_8$	$0.79 = a_9$

To proceed further colour matches are represented by equations. These equations are written in the form:

$$C(C) \equiv R(R) + G(G) + B(B)$$

Where the equivalence sign means 'matches', the letters C, R, G and B represent the amount of colour used and the letters in brackets are labels to which the amounts refer

The chromaticity co-ordinates, a_1 to a_9, can be used to represent the amounts of XYZ needed in colour matching equations to match amounts of RGB, as follows:

$$k_1(R) \equiv a_1(X) + a_2(Y) + a_3(Z)$$
$$k_2(G) \equiv a_4(X) + a_5(Y) + a_6(Z)$$
$$k_3(B) \equiv a_7(X) + a_8(Y) + a_9(Z)$$

By treating this set of three equations as a matrix the subject of them can be changed to:

$$1.0(X) \equiv c_1k_1(R) + c_2k_2(G) + c_3k_3(B)$$
$$1.0(Y) \equiv c_4k_1(R) + c_5k_2(G) + c_6k_3(B)$$
$$1.0(Z) \equiv c_7k_1(R) + c_8k_2(G) + c_9k_3(B)$$

Where the values of c are gained by inverting the 3x3 matrix containing the values of a:

$$c_1 = 2.06 \; c_2 = -1.14 \; c_3 = 0.08$$
$$c_4 = -0.94 \; c_5 = 2.21 \; c_6 = -0.27$$
$$c_7 = -0.32 \; c_8 = 0.05 \; c_9 = 1.27$$

The colour matching equations for the chosen reference white, to which the amounts of RGB previously stated produce, are:

$$k_4(W) \equiv H_1(R) + H_2(G) + H_3(B)$$
$$k_4(W) \equiv J_1(X) + J_2(Y) + J_3(Z).$$

Where H_1, H_2, H_3 are the amounts of (R), (G), (B), needed to match the white, and J_1, J_2, J_3 are proportional to the x, y, z chromaticity co-ordinates of the white. J_2 is equal to the luminance factor of the white.

Substituting for X, Y and Z in the above equation gives:

$$k_4(W) \equiv J_1c_1k_1(R) + J_1c_2k_2(G) + J_1c_3k_3(G) + J_2c_4k_1(R) + J_2c_5k_2(G) + J_2c_6k_3(G) + J_3c_7k_1(R) + J_3c_8k_2(G) + J_3c_9k_3(G)$$

This enables the derivation of H_1, H_2 and H_3:

$$H_1 = (J_1c_1 + J_1c_4 + J_1c_7)\, k_1$$
$$H_2 = (J_2c_2 + J_2c_5 + J_2c_8)\, k_2$$
$$H_3 = (J_3c_3 + J_3c_6 + J_3c_9)\, k_3$$

And hence the equations for k_1, k_2 and k_3:

$$k_1 = H_1 / (J_1c_1 + J_1c_4 + J_1c_7)$$
$$k_2 = H_2 / (J_2c_2 + J_2c_5 + J_2c_8)$$
$$k_3 = H_3 / (J_3c_3 + J_3c_6 + J_3c_9)$$

Using D_{65} as the standard white gives J_1=95.04, J_2=100, J_3=108.89 and H_1=H_2=H_3=100. The above equations can now be resolved.

$$k_1 = 1.49$$
$$k_2 = 0.85$$
$$k_3 = 0.84$$

Returning to the earlier set of equations:

$$k_1(R) \equiv a_1(X) + a_2(Y) + a_3(Z)$$
$$k_2(G) \equiv a_4(X) + a_5(Y) + a_6(Z)$$
$$k_3(B) \equiv a_7(X) + a_8(Y) + a_9(Z)$$

They can be rearranged to give:

$$1.0(R) = a_1/k_1(X) + a_2/k_1(Y) + a_3/k_1(Z)$$
$$1.0(G) = a_4/k_2(X) + a_5/k_2(Y) + a_6/k_2(Z)$$
$$1.0(B) = a_7/k_3(X) + a_8/k_3(Y) + a_9/k_3(Z)$$

By substituting this into the generic colour matching equation:

$$C(C) \equiv Ra_1/k_1(X) + Ra_2/k_1(Y) + Ra_3/k_1(Z) + Ga_4/k_2(X) + Ga_5/k_2(Y) + Ga_6/k_2(Z) + Ba_7/k_3(X) + Ga_8/k_3(Y) + Ba_9/k_3(Z)$$

The RGB to XYZ transform equations can then be derived:

$$X = (a_1/k_1)R + (a_4/k_2)G + (a_7/k_3)B$$
$$Y = (a_2/k_1)R + (a_5/k_2)G + (a_8/k_3)B$$
$$Z = (a_3/k_1)R + (a_6/k_2)G + (a_9/k_3)B$$

The values for the constants can be inserted, hence:

$$X = 0.64/1.49R + 0.29/0.85G + 0.15/0.84B$$
$$Y = 0.33/1.49R + 0.60/0.85G + 0.06/0.84B$$
$$Z = 0.03/1.49R + 0.11/0.85G + 0.79/0.84B$$

A similar method can be used to find the reverse transform:

$$R = 2.06*1.49X - 0.94*1.49Y - 0.32*1.49Z$$
$$G = -1.14*0.85X + 2.21*0.85Y + 0.05*0.85Z$$
$$B = 0.08*0.84X - 0.27*0.84Y + 1.27*0.84Z$$

4.3. Transform from XYZ to Lab

The conversion from XYZ to Lab is more routine:

$$L = 116(Y/Y_n)^{1/3}-16$$
$$a = 500[(X/X_n)^{1/3}-(Y/Y_n)^{1/3}]$$
$$b = 200[(Y/Y_n)^{1/3}-(Z/Z_n)^{1/3}]$$

This is only complicated when looking at dark areas of a scene. If any of the ratios X/X_n, Y/Y_n or Z/Z_n is equal to or less than 0.00856, it is replaced in the above formula by:

$$7.787F + 16/116$$

Where F is X/X_n, Y/Y_n or Z/Z_n as appropriate

These transforms can be performed in the reverse direction to turn Lab into XYZ:

$$Y = Y_n\left(\frac{L+16}{116}\right)^3$$

Or if $Y/Y_n \le 0.008856$

$$Y = Y_n\left(\frac{\left(\frac{L+16}{116}\right)^3 - \frac{16}{116}}{7.787}\right)$$

$$X = X_n\left(\frac{a}{500} + \left(\frac{Y}{Y_n}\right)^{\frac{1}{3}}\right)^3$$

Or if $X/X_n \le 0.008856$

$$X = X_n\left(\frac{\left(\frac{a}{500} + \left(\frac{Y}{Y_n}\right)^{\frac{1}{3}}\right)^3 - \frac{16}{116}}{7.787}\right)$$

$$Z = Z_n\left(\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - \frac{b}{200}\right)^3$$

or if $Z/Z_n \le 0.008856$

$$Z = Z_n\left(\frac{\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - \frac{b}{200}\right)^3 - \frac{16}{116}}{7.787}\right)$$

If $Y/Y_n \le 0.008856$ then it should be replaced in the above four equations by $7.787\ Y/Y_n + 16/116$.

These equations are best combined in the form of a short program to perform the Lab to rgb and rgb to Lab conversions more easily.

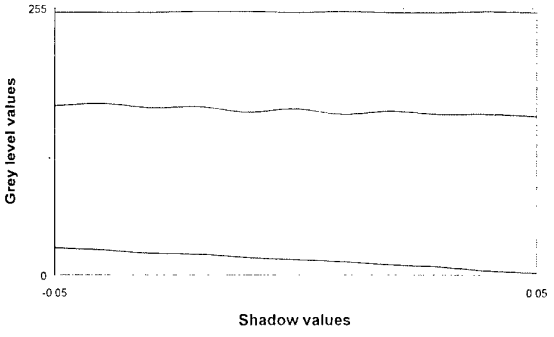
5. CALIBRATION

With these conversions accomplished, it is possible to compare the input and output of the scanning and printing processes. The next stage is to model the effects of the scanner and printer control functions. This will enable the prediction of the settings required to retain accurate colour registration through the system.

Firstly the effect of each of the scanner controls needs to be measured. The values obtained from a test slide scanned on the default settings are recorded. Then the controls are individually changed through small increments and the new values recorded. This data can be examined to determine the type of function that each control is.

Figure 2 Shows how the shadow control changes the grey level values.

Through inspection it can be seen that as the value of the shadow control increases the grey level values decrease by an amount proportional to the difference between the maximum value of 255 and the original grey level. This indicates that the



shadow control is the following function type:

$$new \approx 255 - \left(\frac{255 - old}{1 - shadow}\right)$$

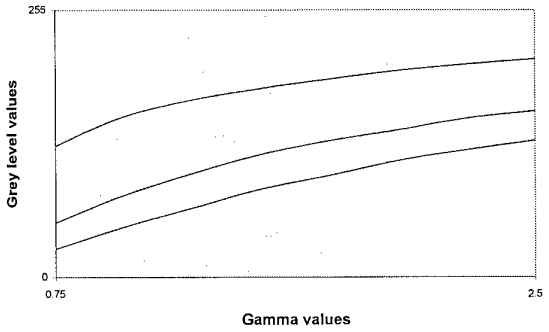


Figure 3 Shows how the gamma control changes the grey level values.

Similarly, it can be seen that as the value of the gamma control increases the grey level increases by a proportionally smaller amount. This indicates that the gamma control is the following function type:

$$new \approx old^{(1/gamma)} * 255 / 255^{(1/gamma)}$$

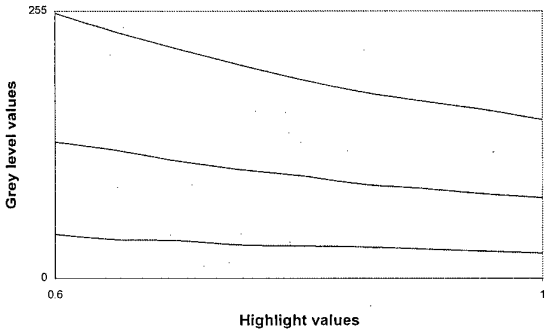


Figure 4 Shows how the highlight control changes the grey level values.

Again, it can be seen that as the value of the highlight control decreases the grey level increases proportionally with its inverse value. This indicates that the highlight control is the following function type:

$$new \approx old / highlight$$

The functions for the individual controls then need to be integrated into a single function to mimic their interaction.

$$new \approx \left(\frac{255 - \left(\frac{255 - old}{1 - shadow} \right)}{highlight} \right)^{\left(\frac{1}{gamma} \right)} * \frac{255}{255^{\frac{1}{gamma}}}$$

This is still only an approximation, requiring the addition of constants of proportionality to enable it to replicate the function of the calibration controls accurately. The values of these constants are best determined by fitting the function to a set of data from a variety of control settings. As this function is likely to require truncating at either end of the 0 – 255 scale at various stages, it will probably take the form of a small program.

This creates another problem, in that a number of data-fitting software packages that allow user defined functions will not accept conditional statements. It is therefore necessary to write a program to fit the function to the data. An adapted version of the Amoeba minimisation routine (Press et al., 1986) has proved satisfactory for this purpose.

With the fitting complete, the variables become constants and the function can be used to predict the setting of the controls.

If the scanner has been calibrated successfully, the printer calibration should be a single set up for a suitable type of film, simply tuning the film and printer to the computer output. Small differences between the projected slide and the scanned image can be corrected using the printer controls.

The modelling of the printer controls follows the same steps as the scanner: establishment of the type of function and then fitting it to a set of sample data using the same methods.

It is important to note the importance of film type at this point. All films used in the printer are likely to achieve the required extremes of scale, white and black, however the response of the film at intermediate levels should also be close to that required. A slow film will make most of the image too dark and a fast film will make most of the image too bright. With this particular printer, a 200ASA film is used to start close to the required response.

The system should now be capable of reproducing a slide perfectly. However, by using the photometric properties of objects in the scene, rather than those of the projected slide, it should be possible to produce an image that has similar colours and contrast levels to the original scene, removing the changes introduced by the photographic process that recorded it.

6. CONCLUSION

Between the scanning and printing phases, this system can be used in the creation of imagery for assessment. It is at this point that targets can be inserted into the scene. It should be obvious that great care should be taken to ensure that colour registration between the target and background is carefully controlled. It is essential that any target inserted should be as appropriate as possible in size, colour and contrast terms. To this end, capture of the base scene will need to be accompanied by measurement of the dimensions of key objects and the spectral distribution of the illumination when the imagery is taken. If a target is obtained photographically, its true reflectance should be recorded as well as the spectral distribution of the illumination. This will allow the target to be tuned to the background scene and inserted using an image manipulation software package. These details will also be required if a 3D modelling package is used to provide the target.

This type of image manipulation facility has the potential to bypass the field trial phase of CCD assessment, thus saving time and money. It could also be used to assess CCD more thoroughly by using a variety of background scenes or the same scene at different times of the year. Targets that do not exist in the real world could be assessed, targets being created from CAD models. It could be used to determine the effectiveness of potential CCD measures in areas that are not readily accessible for field trials. Another use may be to remove some of the effects on colour and contrast due to the nature of the photographic media and its projection. Images corrected in this way might appear more realistic than a standard slide.

Overall, this system adds a new level of flexibility and completeness to photosimulation.

7. REFERENCES

Ashforth, M. and Collins, J.H. (1991) Determination of detection range by analysis of recorded imagery (Technical Memorandum SCRDE/91/6), Colchester, United Kingdom: Science and Technology Division, Defence Clothing and Textiles Agency.

Hunt, R.W.G. (1987) Measuring Colour. Ellis Horwood Ltd.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) Numerical Recipes. Cambridge University Press.

A PHYSICS BASED BROADBAND SCENE SIMULATION TOOL FOR CCD ASSESSMENT

Dr I. R. Moorhead[+], Mrs M. A. Gilmore[*], Mr D. Oxford[#], Dr D Filbee[♦], Colin Stroud[♦], G Hutchings[♦], A Kirk[♦]

[+] Protection & Performance Dept, Centre For Human Sciences, e-mail: I_Moorhead@dera.gov.uk

[*] Airborne signatures and IRCM, Weapons Systems Sector, e-mail: MAGilmore@dera.gov.uk

[#] Sensors & Avionic Systems Dept, Sensors & Processing Sector, e-mail: deoxford@dera.gov.uk.,

Defence Evaluation and Research Agency, Farnborough, Hants GU14 OLX

[♦] Hunting Engineering Ltd (HEL), Reddings Wood, Ampthill, Bedford, MK45 2HD, UK

e-mail: drf@hunting2.demon.co.uk

1. SUMMARY

Assessment of Camouflage, Concealment and Deception (CCD) methodologies is a non trivial problem; conventionally the only method has been to carry out field trials, which are both expensive and subject to the vagaries of the weather. In recent years computing power has increased, such that there are now many research programmes using synthetic environments for CCD assessments. Such an approach is attractive; the user has complete control over the environment parameters and many more scenarios can be investigated.

The UK Defence Evaluation and Research Agency is currently developing a synthetic scene generation tool for assessing the effectiveness of air vehicle camouflage schemes. The software is sufficiently flexible to allow it to be used in a broader range of applications, including full CCD assessment. The synthetic scene simulation system (CAMEO-SIM) has been developed, as an extensible system, to provide imagery within the 0.4 - 14 micron spectral band with as high a physical fidelity as possible. It consists of a scene design tool, an image generator, which incorporates both radiosity and ray-tracing processes, and an experimental trials tool. The scene design tool allows the user to develop a three-dimensional representation of the scenario of interest from a fixed view-point. Target(s) of interest can be placed anywhere within this 3-D representation and may be either static or moving. Different illumination conditions and effects of the atmosphere can be modelled together with directional reflectance effects. The user has complete control over the level of fidelity of the final image. The output from the rendering tool is a sequence of radiance maps which may be used by sensor models, or for experimental trials in which observers carry out target acquisition tasks. The software also maintains an audit trail of all data used to generate a particular image, both in terms of material properties used and the rendering options chosen.

Keywords: Scene Simulation, CCD Assessment, Camouflage, Concealment & Deception

2. INTRODUCTION

All camouflage is a compromise. It is required to match different backgrounds, in different wavebands and at different times of the year. The compromises made in the past were determined by subjective assessment of the visibility of a military asset when viewed against some relevant background. Typically, this assessment was carried out in the visible band only. Two technologies are driving the need for quantitative CCD assessment. Firstly, sensors now operate throughout a large part of the electromagnetic spectrum and it is likely that future sensors will place even greater demands on camouflage design by requiring exact spectral matches. Secondly, new techniques and materials offer the potential of increased effectiveness of camouflage against sensor threats. Cost-effective and quantitatively correct assessment of these techniques and materials is essential for future system survivability.

Synthetic scene generation offers a viable alternative to field trials for the quantitative evaluation of camouflage. We are developing a physics based, broadband, scene simulation toolset called CAMEO-SIM that enables the quantitative evaluation of both current and future camouflage. The same toolset may also be used to assess concealment and deception methodologies.

3. STRUCTURE OF THE PAPER

Section 4, provides an overview of the components that make up the CAMEO-SIM toolset. Section 5 reviews the verification tests that have been carried out to date, section 6 describes the validation programme and section 7 presents conclusions.

4. OVERVIEW OF CAMEO-SIM

The goal of the CAMEO-SIM system is to produce synthetic, high resolution, physically accurate radiance images of target vehicles in operational scenarios, at any wavelength between 0.4 and 14 microns. The main components of the system are shown schematically in Figure 1. These are described in detail elsewhere [1]. The software was developed with a scaleable rendering kernel in which imagery can be produced at different fidelities and frame rates depending on the image application and wavelength of operation.

The lowest fidelity mode is real-time and can be used to develop and preview the scene before it is passed to a scaleable two-pass rendering kernel that can produce high quality image streams using BRDF capable radiosity and ray-tracing algorithms. The first pass computation models the radiative transfer between extended surfaces; ie., it models 'soft' shadow effects. The second pass is a ray tracer to model the effect of point sources - it models the 'hard' shadows. The images produced from these algorithms are then used in the camouflage assessment process. The user has detailed control of the parameters used in the calculations.

The rendering equations solved during real-time operations are termed 'local' rendering equations. The term 'local' is used to acknowledge the fact that the radiative interactions in the scene are predefined. In order to run in real-time, approximations are made during the rendering process. These include:

1. the directionality and bi-directionality of the optical properties of materials in the scene are approximated using scalar diffuse optical properties, perhaps with a simple specular parameter added to the reflectivity.
2. global illumination effects are not accounted for.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

3. geometric occlusion of point and extended sources of radiation is ignored.
4. spectral integration of the optical properties with the atmospherics is reduced to a multiplication of the in-band optical property with the in-band atmospheric term – such an approximation is only valid for spectrally grey materials.
5. the parameterisation of the atmospheric terms is simplified - for example 3D variations in the path radiance and transmittance are ignored.
6. polygon budgets are restricted - which means that the structure of complex objects such as vehicles or trees has to be simplified.

To assess the effectiveness of camouflage schemes, especially on helicopters flying at low level using trees as natural cover, these real-time approximations can impact critically on the final computed air vehicle contrast. Therefore higher fidelity, non real-time computations must be used. The CAMEO-SIM renderers do not make the approximations listed above. The output from the high fidelity rendering kernel is a sequence of floating point, 2-D radiance maps for a given waveband or bands. These may then be played back at real-time rates for different applications.

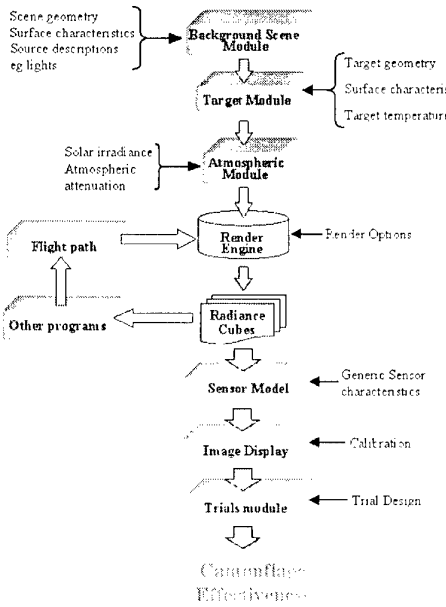


Figure 1 Block diagram of CAMEO-SIM components illustrating the data flow through the different processes.

5. ANALYTICAL VERIFICATION TESTS

CAMEO-SIM Version 1.0 is complete and is now undergoing verification and validation. A range of verification tests has been developed that exercise different elements of the high fidelity rendering equations implemented within CAMEO-SIM. All the tests have analytic solutions. Table 1 summarises the test results. A detailed description of each test is given in the following sections.

Test	Expected result	Calculated
Blackbody radiance	Blackbody Radiance = 42.89 (8-12.5 micron band)	42.89
Contrast in isothermal environment	Centre pixel radiance = 35.23 (8-12.5 micron band)	35.23
Shadowing and Blocking	a. Radiance of irradiated area = 5.1768 b. Radiance of blocked area = 0.0 c. Radiance of shadowed area = 0.0 (3-5 micron band)	a. 5.1768 b. 0.0 c. 0.0
Spectral calculations	Centre pixel radiance (3-5 micron band) = 1.49	1.49
Radiometric calculation of lighting effects	Radiance variation: Centre : 0.31831 edge : 0.0094248	Centre: 0.31806 edge: 0.0094239
Directional emission	Slope of radiance along centreline = 60.01 W m^{-2} $pixel^{-1}$	59.932 W m^{-2} $pixel^{-1}$
Multiple material Assignment on a texture	Blackbody radiance = 8.975 Grey body radiance =4.4875 (3-5 micron band)	Blackbody radiance = 8.975 Grey body radiance = 4.4875
Bi-directional reflectivity	Illuminated pixel radiance = 2.3	2.3
Small target rendering	Integrated facet radiant intensity = 1.806 W sr^{-1} (3-5 micron band)	1.813 W sr^{-1}

Table 1 Summary of validation test results. All values are W m^{-2} sr^{-1} unless otherwise stated.

5.1. Blackbody radiance tests

The purpose of this test was to ensure that the blackbody radiance is calculated correctly. A one metre square uniformly textured facet was created and the temperature of the facet set to a known value. The line of sight of the observer was centered and perpendicular to the facet. The radiance of a perfect blackbody was calculated and compared with the value computed within CAMEO-SIM.

5.2. Contrast in an isothermal environment

The purpose of this test is to ensure that the correct radiance contrast is predicted for isothermal vacuum, radiometric environments.

The skyshine radiance terms are set to constant values. A one metre square surface is defined to be a perfect diffuse reflector and the line of sight of the observer is centered and perpendicular to the facet. The radiance of the square is calculated and compared with the value computed within CAMEO-SIM.

5.3. Calculation of shadowing and blocking

Blocking is the rendering process that ensures parts of the object that are not visible to the observer due to obstruction by another part are correctly accounted for. Shadowing is the rendering process that ensures parts of the object do not reflect the point sources if they are obscured from it by other parts. This test has been designed to ensure that the blocking and shadowing algorithms are working accurately. The geometry for this test is shown in Figure 2 which shows two square plates with the lower plate 100% diffuse reflecting and the top

plate black and at zero Kelvin. The observer and sun are at 45 degrees to the geometry. The radiance of the illuminated pixels in the image is:

$$N = Q\,\rho/\pi \qquad (1)$$

where:

N is the radiance in W m^{-2} sr^{-1}

Q is the normal incident irradiance W m^{-2} = solar irradiance x cosine of incidence angle

ρ is the diffuse reflectance of the lower plate

The solar irradiance is set to a fixed value. The radiance of the shadowed, blocked and irradiated areas is calculated and compared with the values computed within CAMEO-SIM.

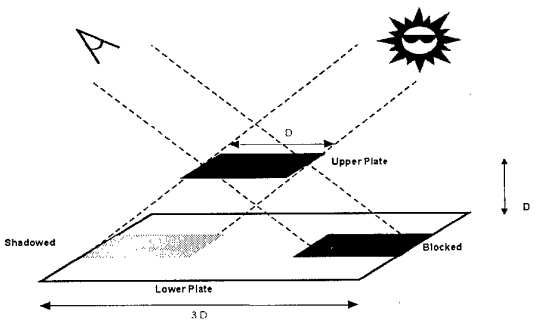


Figure 2 Diagram showing the geometry used to verify the blocking and shadowing computations.

5.4. Spectral Calculations

The purpose of this test specification was to verify that the spectral integrations were being calculated accurately. To test this a defined solar spectral irradiance was used to illuminate an artificial spectral material being observed with a spectrally selective sensor.

The spectral variation in the material properties, the light intensity and the sensor response is defined. For the general case, the in-band reflected radiance between the upper and lower wavelengths is given by:

$$N_{\Delta\lambda} = \int_{\lambda_1}^{\lambda_2} \frac{J(\lambda)}{s^2} \cos\theta_i . \theta(\lambda) . \rho''(\lambda) . d\lambda \qquad (2)$$

where:

$N_{\delta\lambda}$ = in band radiance

θ_i = incidence angle between source and reflector

$J(\lambda)$= source intensity (W sr^{-1})

$\theta(\lambda)$ = sensor spectral response

$\rho''(\lambda)$ = spectral bi-directional reflectivity

s = distance to the source

5.5. Radiometric calculation of lighting effects

The purpose of this test was to verify that the radiometric effects of light sources are being accurately represented. The geometry of the test is shown in Figure 3a and a plot of computed (red line) and rendered radiance is shown in Figure 3b, together with the difference between the computed and rendered radiance. It must be noted the analytical solution assumes radiant intensity is at the pixel's centre, but the image's radiant intensity is super sampled across a pixel. This will introduce a small difference to the analytical solution.

5.6. Directional emission of uniformly textured and heated spheres

This test verifies that the second pass renderer is accounting for the directional emissivity correctly when the object is nominated as having directional optical properties.

Two uniformly textured spheres of 2m diameter are set to a known temperature. For one of the spheres, the vertex normals are equal to the facet normal and for the other, an appropriate angle is chosen for generating the vertex normals. Therefore, in the test both flat faceting and vertex normal interpolation in the second-pass renderer are tested. The variation in pixel radiance from the centre of the sphere to the outside edge should vary linearly (for the vertex normal interpolated sphere, and approximated with a stepped variation for the flat facet sphere.

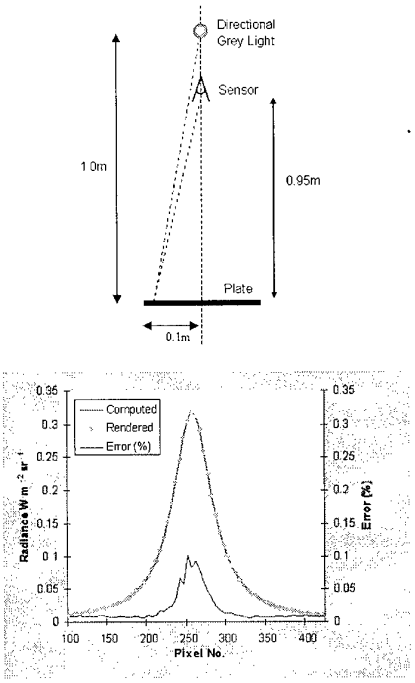


Figure 3 a) Lighting effects geometry. b) graph of computed and predicted radiance as a function of pixel position.

5.7. Textured heated billboard for testing multiple material assignments on a texture.

The purpose of this test is to ensure that textures that have been classified using multiple material associations and transparency are interpreted properly by CAMEO-SIM. To test this aspect a heated non-uniformly textured billboard with a transparent section is rendered. A 256 x 256 texture image containing two rectangles and a transparent region is created. One rectangle is classified as a blackbody perfect diffuser and

set to a known temperature. The other rectangle is set to be a greybody perfect diffuser at the same temperature. A typical image expected from this test is shown in Figure 4.

5.8. Bi-directional reflectivity of uniformly textured and heated spheres.

The purpose of this test was to verify that CAMEO-SIM is interpreting the bi-directional reflectance function correctly.

To keep the solution to the BRDF problem analytically tractable a BRDF file was used that represents a grey semi-specular retro-reflecting BRDF such that

$$BRDF = \frac{1}{\cos(\theta)} \qquad \theta \leq 30\,\text{deg}$$

$$BRDF = 0.0 \qquad \theta > 30\,\text{deg}$$

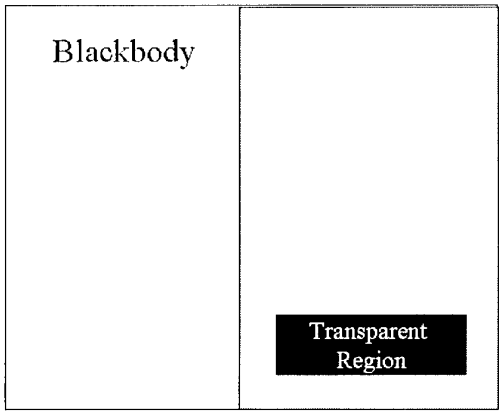


Figure 4 Schematic of the image produced by the textured heated billboard test.

where 0 is the angle of incidence. Two spheres are created: for one sphere the vertex normals are equal to the facet normal and for the other sphere an appropriate angle is chosen for generating the vertex normals. The line of sight of the observer is set to view the spheres from above with the sun position above the observer. A typical image in which the illuminated pixels have a nearly constant radiance across their diameters is shown in Figure 5.

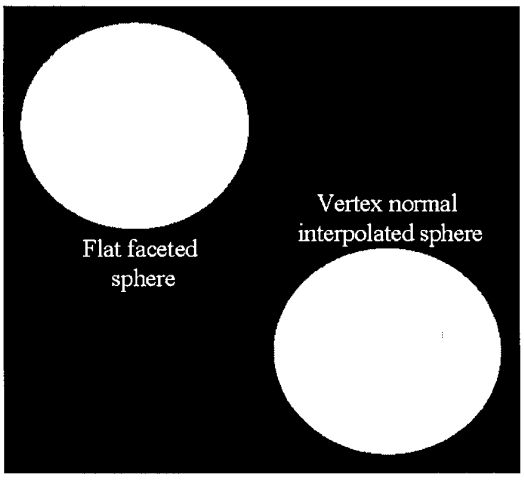


Figure 5 Image resulting from the bi-directional reflectivity test.

5.9. Small target rendering

The purpose of this test was to ensure that CAMEO-SIM is treating small targets to an acceptable accuracy; an essential requirement for simulating potentially sub-pixel targets. To test this requirement an identical sphere to that used in the BRDF test is rendered against a simple uniform background. The geometry of the test case is shown in Figure 6a, and the image formed for this test case should be similar to that shown in Figure 6b. The predicted integrated facet radiant intensity is 1.806 W m^{-2} sr^{-1}. The CAMEO-SIM integrated facet radiant intensity is 1.813 W m^{-2} sr^{-1} - a percentage difference of 0.39%.

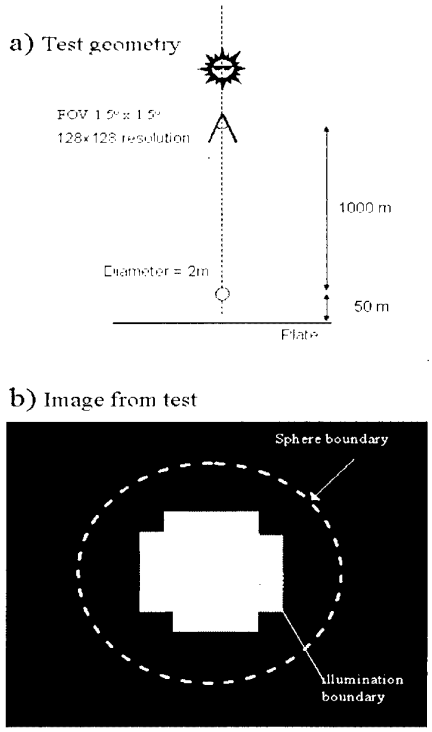


Figure 6 a) Geometry used to carry out the small target test. b) A typical image produced by the test.

5.10. Fit for purpose imagery

The verification tests have shown that CAMEO-SIM is computing the correct values. However rendering times are excessively long if the complete solution is calculated - especially for a complex structured scene. CAMEO-SIM was designed therefore to render images at different levels of fidelity, for different applications - such that the imagery is 'fit for purpose'.

Some example images for a clear day, visible band, generated at different levels of fidelity are shown in Figure 7. These images were created using three-dimensional models of trees constructed from triangular facets. No sensor effects have been added to the images.

Similar images can be created for different wavebands including visible, 3 - 5 micron and 8-12 micron band (Figure 8). A subjective analysis of the above images shows that the significance of different effects varies with waveband and with the weather condition - as expected. For example, on a cloudy day the hard shadows from the sun are not relevant.

Shadows appear to have a more significant effect in images of wavebands < 5 microns than for 8-12 micron band. In the lower wavebands the 'soft' shadow effects produce the three-dimensional effects of shadows on trees which appears to have a large effect on the contrast structure within the image, but this imposes a heavy computational load on the renderer.

Directional reflectance effects give rise to glints and cause more structure to be visible within the object - this is going to have a very significant effect on the spatial contrast structure within the image.

The image in Figure 9 shows the result of differencing and histogram equalising a 'low' fidelity and a 'high' fidelity image. Clearly there are large differences that could contribute significantly to errors in target conspicuity. For an image to be 'fit for purpose' the errors shown in the difference image must be insignificant for that particular application.

When viewed through a sensor the image resolution will be degraded and hence a lower level of fidelity may be acceptable. In addition the 'real world' is inherently variable, so the images only have to be accurate within the limits of the natural variations whilst still capturing the spatial and spectral structure.

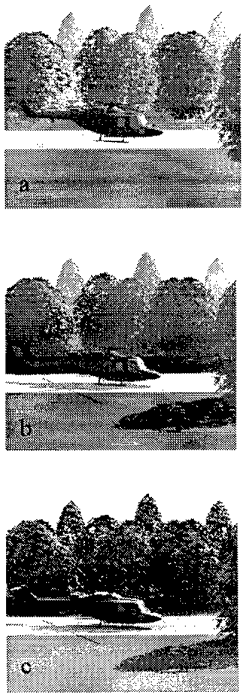


Figure 7 Example images rendered to different levels of fidelity: a) all surfaces diffuse, b) hard shadowing added, c) hard and soft shadowing plus BRDF characteristics applied to the aircraft.

6. VALIDATION

6.1. Issues

The issues surrounding the validation of any piece of simulation software are often complex and CAMEO-SIM is no exception. Furthermore, the fact that CAMEO-SIM aims to physically represent the real world, in many electromagnetic wavebands adds considerably to the difficulties, since we still have neither the basic databases nor the necessary understanding of what constitutes the real world. In addition, since the whole purpose of CAMEO-SIM is to represent scenarios that may not exist or are impossible to document, there may in fact be no equivalent real world. This can be illustrated at the simplest level by considering the geometry and culture that are used to describe a scenario. It is possible to achieve an exact match with the terrain geometry by using detailed map information, but it is impossible to achieve exactly the same geometry for the culture present in that terrain (eg tree structure). This means that validation methods that assume there is some real world database of measurements that can be directly compared with the output from the simulation cannot work. The validation processes that we are using are, as a result, somewhat more abstract and involve three separate approaches. Firstly, to use highly simplified scenarios that can be both synthesised within CAMEO-SIM and measured.

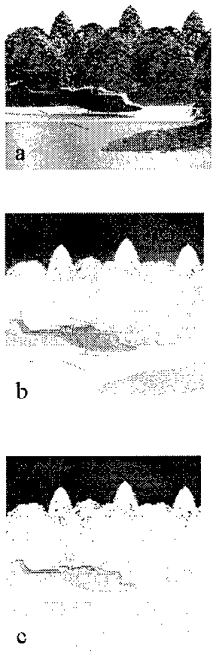


Figure 8 Examples of images generated in different wavebands: a) visible band, b) 3 – 5 μ, c) 8 – 12 μ

Figure 9 Histogram equalised difference between a low fidelity and a high fidelity image with identical geometry.

These move one step away from the basic analytical verification tests described in section 5. Secondly, we are

examining whether the statistics of the imagery produced by CAMEO-SIM are consistent with the statistics of real world images. Finally, since the main use of this tool is for camouflage assessment we will use it to reproduce a real-world trial. This will assess whether the observer performance, using the synthetic imagery CAMEO-SIM generates, corresponds to the performance actually measured in the field. Each of these validation approaches is described in more detail below.

6.2. Simple Imagery Validation

None of the scenarios used in the tests described section 4 were real. The next step therefore is to exercise CAMEO-SIM with imagery that is more realistic. We plan to conduct a series of trials involving imaging a simple object viewed against a uniform background. The object is a metal step-like structure and is shown in Figure 10. The object is sufficiently complex to enable us to exercise both the radiosity and ray tracing processes within the software. Radiation will be reflected from the riser of the step down onto the lower step, the object will generate shadowing, different areas will heat up differentially, etc. It will be placed in the open and viewed by different sensors looking down into the "step" area. Imagery will be gathered in the visible band, MWIR, and LWIR under different conditions. The same scenario will be constructed within CAMEO-SIM using the measured geometry and surface material properties as well as metereological data.

Comparisons will be made between the in-band radiance values measured on the real object under different conditions and the equivalent values calculated within CAMEO-SIM using different rendering fidelities.

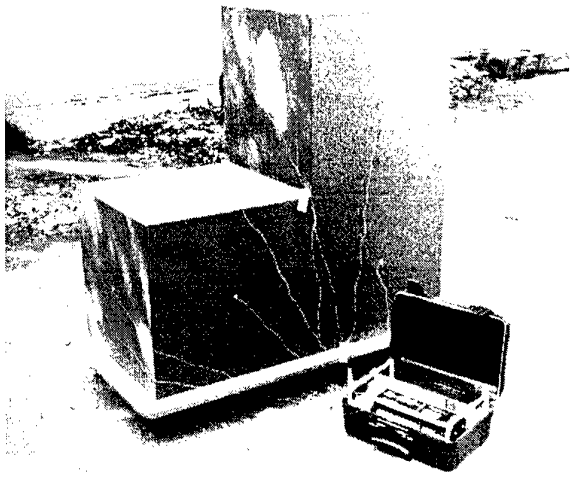


Figure 10 Photograph of the step object to be used in validation trials. The photograph also shows how the object can be instrumented with thermocouples for temperature measurement. (Lighter patches in the image are shadows.)

6.3. Image Statistics Validation

Many image metrics have been proposed [2,3,4,5,6]. Ideally, pixel comparison would be the ultimate method of comparing real and synthetic images. However this would, excessively overspecify the accuracy required, not only because of the natural variations in the real-world, but also because it is often the case that the synthesized imagery does not correspond directly to any part of the real world.

Bivariate metrics which compare similar images, and which are used extensively in the field of image compression (7) cannot be used because it will be impossible to exactly recreate the geometric structure of the real world in the synthetic image - and not necessary. Application of such metrics would indicate that there were large differences between the real and synthetic images, but these differences are not meaningful for this application.

Univariate metrics based on statistics of a single image are more appropriate as an 'image quality' metric. However first order statistics, such as mean and standard deviation are not an appropriate measure for the spatial structure in an image because two totally different images can have the same first order statistics.

Similarly second order statistics, such as the power spectrum, do not completely describe natural images. However, it is clear that if our simulation tool is not even capturing the first order statistics of the real world it has questionable validity. To illustrate this we provide here a comparison we have carried out between the colour characteristics of real world images and equivalent synthetic ones (using a visible-band precursor to CAMEO-SIM called CAM-SIM). A set of real world images obtained by a photographic colorimetric method (8) provided the necessary data for the real world. These were obtained during a trial in 1982 in Southern England. The scenario consisted of open grass fields with scrub and woodland clumps. There were no man-made buildings present in the images although a portion of the image was occupied by a military vehicle. Images were captured on different summer days and at different ranges from 0.5km to 3km. A similar scenario was constructed using CAM-SIM. Colour statistics were collected for the two types of image and the resulting histograms are shown in Figure 11. It is clear that the mean colour, expressed as chromaticity, of the two sets of imagery is similar, but that the distributions and ranges of the synthetic and real imagery are quite different. For certain applications these differences may be critical. Similar comparisons on a range of first and second order statistics are planned for CAMEO-SIM.

Higher order statistics which can capture phase information are likely to be the most appropriate metrics to investigate. However these are complex and hence difficult to interpret and apply. Certain types of Neural networks (Independent Component Analysis networks) naturally capture some of the features of higher order statistics after training. These networks can be trained on real imagery and then used on synthetic images to find if the same characteristics are detected. Therefore it is believed that neural networks would be an appropriate statistical solution for this project.

This work will continue and different metrics used to analyse the synthetic image fidelity so that appropriate images are used for different applications.

6.4. Performance Validation

The third element of the validation programme will involve comparisons of performance. The aim is to reproduce a real trial using imagery generated by CAMEO-SIM.

A flight trial was conducted in 1974 to compare a light grey camouflage scheme with a dark grey /green scheme on UK Royal Air Force strike aircraft. The comparison was made by flying aircraft painted in the two schemes against ground based observers under a variety of meteorological conditions and recording the detection ranges obtained. Observation was

made with the unaided eye and with magnifying sights (X5 and X10). The nature of the trial required the aircraft to fly accurate straight-line tracks of about 20kms on a variety of headings and at low level.

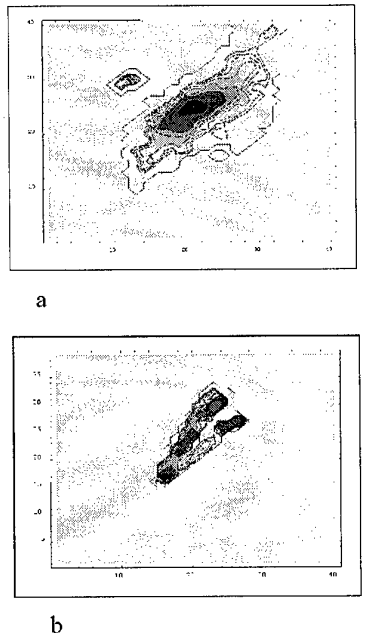


Figure 11 CIE (x,y) chromaticity histograms. a) distribution for real imagery. b) distribution for synthetic imagery.

The following conditions and variables were selected:

1. Observers used unaided viewing, x5 and x10 optical sights.
2. Limited search - the observers knew the approximate height and direction of approach but were not accurately laid–on. Search using the magnifying sights was within the field-of-view along a fixed sight-line
3. The trial was conducted under three types of sky condition – clear sky, 2/8-4/8 broken cloud and uniform overcast. Visibility was at least 10km.
4. Three relative aircraft/observer/sun positions were studies: sun directly behind the observer as he viewed the approaching aircraft, sun at 60° to the approach path shining onto the front of the aircraft and sun at 45° to the approach path shining onto the rear of the aircraft.
5. Approaches were made head-on to the observers and to cross 1000m to left and right of the observers. Approaches began 20km from the observers.
6. Aircraft were 200m above ground level, speed 300 knots.

There are 27 possible different combinations of the above variables, but because sun position is not relevant under a uniformly overcast sky and crossing to left and right was not required when the sun was directly behind observers the actual number of different conditions reduces to18. Each aircraft made 6 approach runs per condition tested – 18 sorties were required to complete the trial.

Both experienced and inexperienced observers were used. Experienced observers had a visual acuity of 6/6 or better with normal colour and binocular vision. Inexperienced observers were tested using a car number plate reading test at a range of 30m.

Cloud conditions generally and in the direction of view were noted. Colour of background sky was estimated by comparison with a set of Munsell colour analysis charts. Background sky luminance was measured with a Weston Master 1V light meter.

Detection ranges were obtained by timing each approach. The observers, with stop watches were given a signal as the aircraft crossed the IP and started their watches, they then stopped their watches individually upon detecting the aircraft. During every approach the aircraft was also timed between two timing marks enabling its ground speed to be accurately determined. This ground speed, the distance between the observers and the IP, and the time to detection were then used to calculate detection range.

The validation exercise will consist of a laboratory simulation of this trial. Clearly attempting to reproduce exactly the same conditions is impossible therefore a much simplified trial design will be used. Visible band image sequences of the scenario containing aircraft in the two different camouflage schemes will be generated using CAMEO-SIM. These will be played back to observers viewing a calibrated colour monitor from a fixed distance (images are rendered to be viewed from a predetermined distance) as shown in Figure 12. Any one sequence will consist of an image of an aircraft "painted" in one of two colours, which will fly along a randomised track towards the observer. The observer task will be to press a mouse button when the aircraft is detected and then to designate the aircraft location on the screen using a cursor.

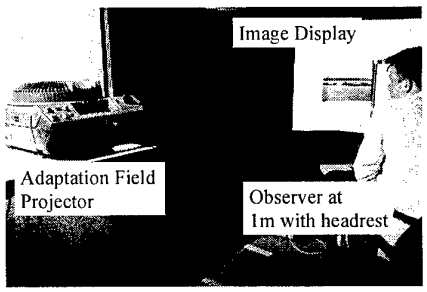


Figure 12 Experimental arrangement. The observer is maintained at a fixed distance from the calibrated display used to present the test imagery. An additional projector is used to provide a controlled adaptation state.

Response time will provide the detection range just as in the real trial. Additionally, this experiment will provide information on accuracy of designation. These data will be compared with the results from the actual trial. Because there are so many differences between the real and the simulated trial conditions only relative performance will be compared. Observers will be screened for colour vision and acuity prior to taking part in the trial. A mean level of adaptation will be maintained by surrounding the display with a white screen if necessary.

7. CONCLUSIONS

CAMEO-SIM can generate imagery between 0.4 and 14 microns to different levels of fidelity, to allow a trade-off between accuracy and rendering time. The software has undergone a range of verification tests to show that the correct values are computed. A programme of validation is now under way to ensure that meaningful results are obtained using the software tools. This programme will address three different aspects of the synthetic imagery – statistics, real-world comparisons and performance prediction. Research work is in progress to quantify the significance of different effects such as shadows, in different wavebands. The functionality of CAMEO-SIM will be extended as part of an ongoing research programme to include multi-processor rendering capability, and simple multi-spectral image display, moving sensors and thermal shadows.

8. ACKNOWLEDGMENTS

This work was sponsored within the UK MoD Applied Research and Corporate Research Programmes.

9. REFERENCES

1. Oxford D., et al. "CAMEO-SIM: A physically accurate broadband EO scene generation system for the assessment of air vehicle camouflage schemes", *Ground Target Modelling and Validation Conference* 1998.
2. Lukas F X J, Budrikis Z L, "Picture quality prediction based on a visual model", *IEEE Trans Comm* 30, pp. 1679-1692, 1982.
3. Nill N B, Bouzas B H, "Objective image quality measure derived from digital image power spectra", *Opt Eng* 31, pp. 813-825, 1992.
4. Fuhrmann D R, Baro J A, Cox Jnr J R, "Experimental evaluation of psychophysical distortion metrics for JPEG encoded images", *J Electronic Imaging* 4, pp. 397 – 406, 1995.
5. Ballard Jr J., Sabol B., Melton Jr R E. "Sensor-based validation of synthetic thermal scenes - how close is good enough?", *Targets and Backgrounds: Characterisation and Representation IV,* Proc. SPIE 3375, pp. 304 - 312, 1998.
6. Brown S., Schott J., Raquero R. "Critical image formation parameters in thermal hyperspectral image simulations", *Targets and Backgrounds: Characterisation and Representation IV.* Proc. SPIE 3375, pp. 313 - 323, 1998.
7. Watson A. B., "Towards a perceptual video quality metric", *Human Vision & Electronic Imaging III,* Rogowitz B.E., Pappas T.N., Eds., Proc. SPIE 3299, pp. 139 – 147, 1998.
8. Burton G. J., Moorhead I.R, "Color and spatial structure in natural scenes", *Applied Optics 26,* pp. 157 – 170, 1987.

AN INVESTIGATION INTO THE APPLICABILITY OF COMPUTER-SYNTHESISED IMAGERY FOR THE EVALUATION OF TARGET DETECTABILITY

M. Ashforth
Defence Clothing & Textiles Agency
Science & Technology Division
Flagstaff Road
Colchester
Essex
CO2 7SS
United Kingdom
E-mail: mlashforth@dcta.demon.co.uk

1. SUMMARY

In the course of an earlier study of the influences on an observer's performance in target detectability assessments, the statistical analysis of the data suggested that there was a difference in the influences at work on an observer between the detection of targets in a real scene and the detection of targets in computer-generated (synthetic) images. Since synthetic imagery is increasingly used in this field, this is an important result. The work described in this report is a further analysis of the original data with the aim of studying more closely this difference. Analysis showed that there is indeed a marked difference between the influence of the observers' visual acuity on their performance in the two types of detection task. The reason is that there is less detailed clutter in synthetic images, which alleviates much of the decision-making an observer has to undergo in detecting a target in a real-scene image. In the synthetic case, the target is either seen or not seen and there is much less uncertainty. This uncertainty, which attends real target detection, swamps any measurable influences on an observer's relative performance in the real-scene case. The conclusion is that computer-generated images used for the evaluation of low-contrast target detection should contain much more clutter detail than at present.

Keywords: Target detection, camouflage evaluation, observer tests, visual acuity, synthetic imagery, visual perception.

2. INTRODUCTION

Evaluation of the effectiveness of camouflage, or, more generally, the measurement of the detectability of low-contrast targets in a cluttered environment, is not a trivial task. Although there are models of human perception, they are at present limited in their applicability, and the case of low-contrast targets in a cluttered environment is the most difficult. Many unquantifiable influences are at work in a human search for inconspicuous targets.

For this reason, the NATO camouflage research community has always relied on the use of numbers of human observers in their evaluation of camouflage effectiveness. This has usually involved photosimulation tests (ref. 1), whereby observers are shown projected photographic images within which a target is concealed. The simulated range at which the target is detected becomes the variable to be tested in the subsequent statistical analysis, whereby individual camouflage measures can be evaluated and compared. Despite the various problems and inadequacies of the test (ref. 2), this remains the most reliable method of camouflage evaluation.

In recent years, computers have made it easy to construct images of targets that do not exist, such as new vehicles in development, or to construct images which are less variable than are real scenes, so that one parameter at a time (e.g. gloss) can be varied, to evaluate its effect on target detectability. These possibilities offer the prospect of an improvement in the method of photosimulation by removing the variability found in real imagery, such as that caused by variations in imaging position, natural illumination, and so on, and also by allowing measurement of the effect of otherwise minor influences on target detectability.

Implicit in the use of computer-generated images in this way is the assumption that the search task for the human observers is the same as for a real scene. Therefore an analysis of observers' performance on computer-generated imagery should show a correlation with their performance on real imagery. An opportunity to test this hypothesis arose during a photosimulation exercise held at the Defence Clothing and Textiles Agency (DCTA) Science and Technology Division (S&TD), in Colchester, United Kingdom, recently.

3. DESIGN OF THE PHOTOSIMULATION TEST

The photosimulation test was set up primarily to evaluate developmental camouflage measures within specific projects, such as for helicopters; for hot, arid environments; and so on. The opportunity was taken to make measurements of other observer-specific attributes that may affect the performance of each observer relative to the pool of observers. It had been hoped that this would enable any quantifiable influence on observer performance to be accounted for, and thereby limit spread in the detection data generated in the photosimulation test.

Some of the imagery used in the test was computer-generated. Although it was not considered at the design stage, this meant that the test also lent itself to the analysis of any difference between real and computer-generated imagery in terms of the dependence of observer performance on any of the measured attributes.

The choice of attributes to include was restricted to those which were intuitively likely to influence observer response and were easy to measure. A brief questionnaire was designed to record details of the observers' age, rank, relevant training, and their normal job within the unit. Tests were devised, with advice from a local optometrist, to measure visual acuity with a Snellen Chart and colour perception with a series of Ishihara Colour Plates.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

Past experience had suggested that some observers were consistently "good" or "bad" in their ability to detect targets in the recorded image. There had been evidence (Annex D of ref. 3) to suggest that observer ability could be accounted for by adjusting the raw data according to how well an observer performed relative to the other observers, and that spread could be reduced as a result.

Observers are familiarised with the nature and procedures of the photosimulation experiment by being shown a pre-test image similar to those that will be shown in the experiment proper. If all observers were shown the same image, and their performance on this image was recorded, it should give a guide to their relative ability. Therefore the final factor to be incorporated was the observer's performance on this pre-test image. This would have the disadvantage that the observers would be learning the procedure at this stage, but the advantage that all observers would see this same image before any of the others, so all saw it under equal terms.

Of the four sets of imagery used in the photosimulation, one consisted of computer-generated imagery. Because the observers were likely to be less familiar with this type of imagery than with real-life photographs, it was decided that the familiarisation image should be computer-generated too.

The photosimulation test was designed so that each observer saw several slides (taken in different locations), some of which had more than one target. This provided data for between 5 and 7 target detections per observer, of which one was in a synthetic image, plus the familiarisation image (also synthetic) that all observers saw. In analysing these detections individually, the assumption is made that they are independent (i.e. one detection does not influence another in the interactive cueing effect). This is not always the case for slides containing more than one target, but no trends were noticed that might have suggested that detections were not independent. Unfortunately 5 to 7 is not a high enough number to conduct a test on the independence of target detections.

4. PRE-TEST DATA

A total of 104 observers were conducted through the trial, all of whom were army personnel from the Colchester Garrison. Their questionnaire responses were coded for entry into an analysis of variance, which would establish how significant each factor was in its contribution to the variance observed in performance. Reference to individuals was made by their Observer Index, which was the number given according to the order in which they were conducted through the whole test. Age was recorded as a whole number of years. Military rank was coded with an integer to represent each level. The military unit to which the observers belonged was recorded, as was the category of job each performed within that unit.

Visual acuity was measured under test conditions and codified in a way suitable to the statistical analysis. Two observers were considered outliers in the visual acuity data. Both of these observers normally wore spectacles, but did not have them available for the test.

Seven of the 104 observers had defective colour vision, and were diagnosed according to the type and degree of deficiency. From a statistical viewpoint, however, so small a sample could not be further subdivided. Colour vision was therefore characterised simply as normal or abnormal. The last category recorded for each observer was the amount of relevant training he had received. All appropriate training was recorded on the questionnaires and was graded by the supervisors with a subjective score out of ten for relevance to the photosimulation task.

5. PHOTOSIMULATION DATA

In order to make detections of different targets comparable, each observer's detection range for a given target was normalised with respect to the mean and standard deviation of all detections made on the same target, as follows.

Z-score for observer against target =

$$\frac{\text{observer's score} - \text{mean score for target}}{\text{sample standard deviation for target}}$$

Thus the Z-score is the amount by which the observer's score exceeds the mean score in units of the standard deviation. A positive Z-score represents a better-than-average result and a negative Z-score represents a worse-than-average result. This removes the differences that exist between the detection difficulty of different targets and allows a comparison to be made of the performance of each observer, relative to the relevant subgroup of observers, i.e. those who detected the same target.

Consistently good observers would be expected to get consistently high Z-scores, so the mean Z-score, averaged over all targets seen by each observer, should be an indication of that observer's ability to detect targets in photosimulation. This, along with the Z-score of the familiarisation slide result, makes two independent measures, designed to be of the same thing.

6. STATISTICAL TESTING

A regression analysis was conducted to determine the correlation between the two sets of Z-scores. A high correlation would confirm that the familiarisation test gives a guide to the ability of the observers. The resulting correlation coefficient was 0.165, which for samples of this size is significant at the 90% confidence level, but no higher. This is not very high and does not give much confidence in the usefulness of the familiarisation slide results as a monitor of observer ability.

Further tests that were conducted to evaluate the effect of the different attributes on observer performance highlighted more differences between the mean Z-scores and the familiarisation Z-scores. These were principally analysis-of-variance (ANOVA) tests, designed to show which of the factors under consideration were contributing to the variance in simulated detection range.

Table 1: ANOVAs on Z-scores (101 Observers)

Factor	df	p (Mean Z)	p (Fam Z)
Age	1	0.432	0.911
Rank	1	0.610	0.637
Unit	3	0.349	0.108
Job	3	0.596	0.021
Colour Vision	1	0.685	0.863
Visual Acuity	1	0.010	0.001
Training	1	0.387	0.269
Error	89		

The three observers who came from training units had to be excluded from the ANOVA because they formed too small a data subgroup. This left 101 observers in the data set. Table 1 shows the results of two separate ANOVAs on the mean Z-scores and the familiarisation Z-scores respectively. This gives a comparison of the relative contribution of each of the

factors to the variance in observer Z-score between the overall mean of the 5 to 7 target detections (the column headed p(Mean Z)) and that for the familiarisation slide (headed p(Fam Z)). The figure in the "df" column gives the number of degrees of freedom for each factor within the analysis. The error term refers to the residual variance. The figures in the "p" columns are the significance levels for each factor: less than 0.05 denotes a significant result, i.e. that the factor has a significant effect on the observers' Z-scores.

Most of the factors included in the analysis have not had a significant influence on either of the sets of Z-scores. In the column for mean Z-scores, only visual acuity has shown a significant effect. It is obvious that in the broadest sense visual acuity will be significant, because if an observer has very poor eyesight, he will not be able to distinguish the targets at all. However, people with very poor eyesight are unlikely to be of interest in a simulation of military target detection and the reason for including this factor was to see if there was an influence even among observers with good eyesight, as mainly used here. There are two observers within the pool who are outliers in the distribution of visual acuity, and they will be exercising a large leverage on the data and its analysis. To check this effect they were removed from the analysis, which was conducted again, exactly as above, but now on the remaining 99 observers. Table 2, below, gives the results of this second analysis.

Table 2: ANOVAs on Z-scores (99 Observers – Visual Acuity Outliers Removed)

Factor	df	p (Mean Z)	p (Fam Z)
Age	1	0.337	0.933
Rank	1	0.526	0.421
Unit	3	0.384	0.060
Job	3	0.667	0.026
Colour Vision	1	0.700	0.983
Visual Acuity	1	0.297	0.001
Training	1	0.400	0.205
Error	87		

Some of the figures in the table have changed, most notably the visual acuity figure for the mean Z-score column, but, importantly, not the visual acuity figure for the familiarisation Z-score column. This is the result that first highlighted the possibility of a difference between the requirements of a search of real imagery and that of synthetic imagery.

Removal of the visual-acuity outliers had the expected effect on the analysis of mean Z-scores, i.e. it removed the apparently significant influence of visual acuity on observer performance (within the narrow spread of visual acuity scores still in the analysis). Remarkably, the same effect was not apparent in the analysis of familiarisation Z-scores; a very significant influence remaining. Note also the other two apparently significant effects; "unit", at 90% confidence; and "job", significant at the 95% confidence level.

If there really is a difference between the requirements of real and synthetic imagery searches, then a closer correlation would be expected between the familiarisation Z-scores and the synthetic-imagery photosimulation Z-scores than that measured earlier between the familiarisation scores and the overall mean ones. This is easily tested. The correlation coefficient for familiarisation Z-scores against the synthetic imagery Z-scores was 0.305, which is significant at the 99.8% confidence level. This is therefore a much more significant correlation than was found with the overall mean results.

Further, if this highly-correlated set of results formed part of the data making up the overall means, then another correlation test should be conducted on the familiarisation Z-scores against the mean of all real-scene Z-scores (that is all except the synthetic-imagery scores). This produces a correlation coefficient of 0.085, which equates to a confidence level of 61%, i.e. not at all significant, or no correlation.

This is a striking result. There is no correlation between the relative performance of observers on the familiarisation slide with that on the 6 real-scene targets, but there is a high correlation with their performance on the other synthetic-image target.

7. DISCUSSION

The statistical work has proved that there is an important difference between target detection from real-scene imagery and detection from computer-generated imagery. This difference has been detected through the relative performance of observers in the target detection task. This infers that some observers are particularly good at detection of targets in real scenes and others are better on synthetic imagery. There must, therefore, be a difference in the demands of each.

The analyses of variance, reported in Section 6, gave a clue when they produced different figures for the significance of the influence of various factors on observers' relative performance. The most notable difference was recorded in the case of visual acuity, which, for the limited spread of acuity found in the 99 observers tested, was not a significant factor in observer performance on real-scene imagery, but was highly significant in the case of synthetic imagery. This implies that detection of targets in synthetic imagery demands good visual acuity, more than does detection of targets in real-scene imagery.

This can be tested specifically, by calculating the correlation coefficient between the visual acuity score and both the mean Z-score for real-scene imagery and the mean Z-score for synthetic imagery. Table 3 shows the results of such an analysis.

Table 3: Correlation of Visual Acuity with Z-Scores

Image Type	Correlation Coeff	p
Real Scene	0.147	0.137
Synthetic	0.381	0.000067

This is an emphatic result. The "p" column gives the probability that the correlation coefficients given could occur by chance if there was no real correlation. It is therefore the significance figure. Within the range covered (by all 104 observers), visual acuity has no significant correlation with the observers' performance in detecting targets in real-scene imagery, even at the 90% confidence level (which would require that $p<0.1$). By the same token, visual acuity is significantly correlated with observer performance in synthetic-imagery target detection at the 99.99% confidence level. Visual acuity would therefore seem to be the main cause of differences in observer performance between the two types of imagery.

There was a suggestion evident in Table 2 that "job" and "unit" may also contribute something to the difference between observers' performance on real and synthetic imagery. One way to test this is to run single analyses of

variance on each data set for each of these two factors. This would produce significance values for each effect. The resulting values are shown below in Table 4.

Table 4: ANOVA for "Job" and "Unit"

Type	p(unit)	p(job)
Real Scene	0.664	0.520
Synthetic	0.555	0.093

The non-significant figures for "unit" suggest that the slightly-significant result in Table 2 (p=0.060) was a rogue. Such a value would be expected by chance roughly once in twenty occasions, so this is quite likely, given the number of tests conducted. The new results above are more reliable than the one in Table 2 because all of the data are used here, whereas some elements had to be removed to do the earlier multiple ANOVA.

Note, however, that there is still a minor difference apparent in the data for "job". There is no significance at all in the effect of "job" on the real-scene data, whereas 0.093, for the synthetic-image data, represents a significant result at the 90% confidence level, though this is not very high and could have occurred by chance.

It would appear that visual acuity is the factor that accounts for almost all of the difference between the demands of real and synthetic imagery in the search for inconspicuous targets. Comparison of the visual appearance of the two types of imagery is necessary in order to attempt to explain this difference.

The reason for the difference is probably that the artificial scene was very homogeneous, using a large number of almost identical-looking trees with a very plain "grass" base. There were few opportunities to be mistaken about the target's whereabouts: it could either be seen or it could not. In real imagery, trees and bushes differ more. There are shady clumps that can look like a camouflaged vehicle. There is much more scope to be mistaken.

In other words, the visual acuity is much more important in synthetic imagery, because there is very little other decision-making to do. When a target is found, it is found with some certainty. In real imagery, there may be many potentially "false" targets, and the observer has to decide how certain he is that he has indeed found a real target. In this case, though visual acuity might be equally important as in the former case, it is swamped by the vagaries of human decision-making in the detection data. Indeed, for real-scene imagery, no factor has been shown in this investigation to have a significant effect on the performance of an observer relative to the pool of observers who detected the same target. The "random error" of the decision process is greater than the effect of any of the individual influences considered here.

8. CONCLUSIONS

An important, and potentially far-reaching, conclusion has emerged from work that was originally designed to evaluate the effect of various potential influences on the performance of observers in the detection of low-contrast targets in a cluttered environment. It is that there is a major difference in the influence of observers' relative performance within the group of observers between target detection in real-scene images and that in computer-generated images.

In essence, the problem is that synthetic images are not sufficiently cluttered to simulate the search task presented by a low-contrast target in a real scene. Computer-generated images are increasingly being used in target detectability studies, on the assumption that such imagery is a sufficiently realistic simulation of real scenes. The work reported here throws doubt on that assumption. In particular it has shown that there is a difference in the demand on observers in the detection task, i.e. that visual acuity is more important in synthetic imagery than it is in real-scene imagery.

The effect of this problem in detectability evaluations will be to introduce a bias that would not show in real-scene work. The observers' visual acuity would influence their own performance. The choice of observers and their distribution across comparative groups would need to be done very carefully with regard to their visual acuity, which would of course need to be tested. Alternatively, by measuring the size of this influence of visual acuity, it could in principle be accounted for by adjusting observers' responses, according to their acuity score.

As computers advance in power, so it should be possible to generate more and more realistic synthetic imagery that would approach the degree of clutter found in photographs of real scenes. This work suggests that that position has probably not yet been reached, and certainly suggests that as much realistic clutter as possible should be included in any synthetic imagery intended for use in an evaluation of the detectability of low-contrast targets.

9. REFERENCES

1. Ashforth, M. and Collins, J.H., "Determination of Detection Range by Analysis Recorded Imagery" (Technical Memorandum SCRDE 91/6, NATO AC/243/CCD/WG(D) 1/91), DCTA – S&T Division, Colchester, UK, 1991.

2. Ashforth, M., "Camouflage Evaluation: Improvements in the Conduct and Analysis of Photosimulation" (DCTA S&TD Research Report 96/02), DCTA – S&T Division, Colchester, UK, April 1996.

3. Ashforth, M., "Evaluation of Handling and Camouflage Properties of Commercially Available Camouflage Nets with Regard to Requirements of SCST 005" (DCTA S&TD Technical Report 94/08), DCTA – S&T Division, Colchester, UK, December 1994.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-MP-45 AC/323(SCI)TP/19	ISBN 92-837-1035-5	UNCLASSIFIED/ UNLIMITED

5. Originator Research and Technology Organization
North Atlantic Treaty Organization
BP 25, 7 rue Ancelle, F-92201 Neuilly-sur-Seine Cedex, France

6. Title
Search and Target Acquisition

7. Presented at/sponsored by
the Workshop of the RTO Systems Concepts and Integration (SCI) Panel held in Utrecht, The Netherlands, 21-23 June 1999.

8. Author(s)/Editor(s)	9. Date
Multiple	March 2000
10. Author's/Editor's Address	**11. Pages**
Multiple	238

12. Distribution Statement There are no restrictions on the distribution of this document. Information about the availability of this and other RTO unclassified publications is given on the back cover.

13. Keywords/Descriptors

Camouflage
Concealment
Target acquisition
Image processing
Target signatures
Models
Simulation
Detection
Performance evaluation
Human factors engineering
Measurement

14. Abstract

This volume contains the Technical Evaluation Report, the Keynote Address, and the 26 unclassified papers, presented at the Workshop on Search and Target Acquisition, that was organised by the Systems Concepts and Integration (SCI) Panel 12 (the former RSG-2), on "Camouflage, Concealment and Deception Evaluation Techniques", and that was held in Utrecht, The Netherlands, from 21-23 June 1999.

The papers presented covered the following headings:
- search performance predictions
- target acquisition mechanisms
- simulation issues

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DIFFUSION DES PUBLICATIONS
RTO NON CLASSIFIEES

L'Organisation pour la recherche et la technologie de l'OTAN (RTO), détient un stock limité de certaines de ses publications récentes, ainsi que de celles de l'ancien AGARD (Groupe consultatif pour la recherche et les réalisations aérospatiales de l'OTAN). Celles-ci pourront éventuellement être obtenues sous forme de copie papier. Pour de plus amples renseignements concernant l'achat de ces ouvrages, adressez-vous par lettre ou par télécopie à l'adresse indiquée ci-dessus. Veuillez ne pas téléphoner.

Des exemplaires supplémentaires peuvent parfois être obtenus auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la RTO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus sur la liste d'envoi de l'un de ces centres.

Les publications de la RTO et de l'AGARD sont en vente auprès des agences de vente indiquées ci-dessous, sous forme de photocopie ou de microfiche. Certains originaux peuvent également être obtenus auprès de CASI.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bunderswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

BELGIQUE
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Directeur - Recherche et développement - Communications et gestion de l'information - DRDCGI 3
Ministère de la Défense nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

ETATS-UNIS
NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GRECE (Correspondant)
Hellenic Ministry of National Defence
Defence Industry Research & Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HONGRIE
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Centro documentazione tecnico-scientifica della Difesa
Via Marsala 104
00185 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

PAYS-BAS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

POLOGNE
Chief of International Cooperation Division
Research & Development Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
VTÚL a PVO Praha /
Air Force Research Institute Prague
Národní informační středisko obranného výzkumu (NISČR)
Mladoboleslavská ul., 197 06 Praha 9

ROYAUME-UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

AGENCES DE VENTE

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
Etats-Unis

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume-Uni

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Les demandes de documents RTO ou AGARD doivent comporter la dénomination "RTO" ou "AGARD" selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications RTO et AGARD figurent dans les journaux suivants:

Scientific and Technical Aerospace Reports (STAR)
STAR peut être consulté en ligne au localisateur de ressources uniformes (URL) suivant:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR est édité par CASI dans le cadre du programme NASA d'information scientifique et technique (STI)
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
Etats-Unis

Government Reports Announcements & Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 2216
Etats-Unis
(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par le Groupe Communication Canada Inc.
(membre de la Corporation St-Joseph)
45, boul. Sacré-Cœur, Hull (Québec), Canada K1A 0S7

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Telefax 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DISTRIBUTION OF UNCLASSIFIED
RTO PUBLICATIONS

NATO's Research and Technology Organization (RTO) holds limited quantities of some of its recent publications and those of the former AGARD (Advisory Group for Aerospace Research & Development of NATO), and these may be available for purchase in hard copy form. For more information, write or send a telefax to the address given above. **Please do not telephone.**

Further copies are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO publications, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your organisation) in their distribution.

RTO and AGARD publications may be purchased from the Sales Agencies listed below, in photocopy or microfiche form. Original copies of some publications may be available from CASI.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Director Research & Development
Communications & Information
Management - DRDCIM 3
Dept of National Defence
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
VTÚL a PVO Praha /
Air Force Research Institute Prague
Národní informační středisko
obranného výzkumu (NISCR)
Mladoboleslavská ul., 197 06 Praha 9

DENMARK
Danish Defence Research
Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (ISP)
29 Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bunderswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)
Hellenic Ministry of National
Defence
Defence Industry Research &
Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HUNGARY
Department for Scientific
Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Centro documentazione
tecnico-scientifica della Difesa
Via Marsala 104
00185 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

NORWAY
Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Chief of International Cooperation
Division
Research & Development
Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

SPAIN
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

TURKEY
Milli Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

UNITED KINGDOM
Defence Research Information
Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
NASA Center for AeroSpace
Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
United States

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
STAR is available on-line at the following uniform resource locator:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR is published by CASI for the NASA Scientific and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
United States

Government Reports Announcements & Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic Database or on CD-ROM)

Printed by Canada Communication Group Inc.
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

ISBN 92-837-1035-5